Unit 6

Normal distributions

Introduction

This unit focuses on normal distributions. Normal distributions are important in statistics as an approximate model for the variability observed in many different contexts. They also arise in situations where a large sample of data is analysed.

The unit begins in Section 1 by introducing the family of normal distributions. These are continuous distributions and, as such, probabilities are found by calculating areas under the graph of the probability density function. You have seen that this is most easily done using the cumulative distribution function. However, there is no convenient formula for the c.d.f. of a normal distribution, so in practice probabilities are found either by referring to printed tables or by using a computer. In Section 2 you will use Minitab to calculate normal distribution probabilities, while in Section 4 you will use printed tables of one particular normal distribution to calculate probabilities associated with any normal distribution. The same remarks apply to quantiles of the normal distribution, which will be treated alongside the c.d.f. and probabilities. In order to use the tables, first some results on linear functions of normal random variables are required, which are presented in Section 3.

In Section 5, normal probability plots are introduced. These plots give a graphical method for investigating whether or not a normal distribution is a good 'fit' for a sample of data. Finally, Section 6 explores modelling the variability which can be observed in large samples of data, and you will see that normal distributions have an important role to play.

1 The family of normal distributions

In this section, the family of normal distributions is introduced. We begin by looking at some datasets.

Example 1 *Heights*

The data in Table 1 (overleaf) are from a study of osteoporosis. This is a disease in which bones lose their strength and become more likely to break. It can be connected with a reduction in height as one grows older. The condition is more prevalent in women than in men. The table includes the heights of a sample of 351 elderly women randomly selected from the community. (Participants in the study may or may not suffer from osteoporosis.) The heights are given to the nearest centimetre; for example, three women among the 351 were 145 cm tall – that is, between 144.5 cm and 145.5 cm.

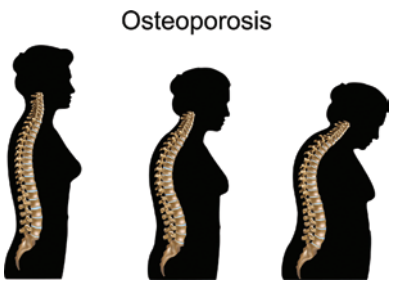


Illustration of how osteoporosis affects height over time

Table 1 Heights of 351 elderly women (cm)

Height	142	143	144	145	146	147	148	149	150	151
Frequency	1	0	0	3	1	4	2	1	6	6
Height	152	153	154	155	156	157	158	159	160	161
Frequency	12	17	11	21	20	20	31	17	21	20
Height	162	163	164	165	166	167	168	169	170	171
Frequency	18	30	17	18	11	7	6	8	11	3
Height	172	173	174	175	176	177	178			
Frequency	0	3	1	0	1	1	2			

(Source: data provided by D.J. Hand, Imperial College London)

A (frequency) histogram summarising these data is shown in Figure 1.

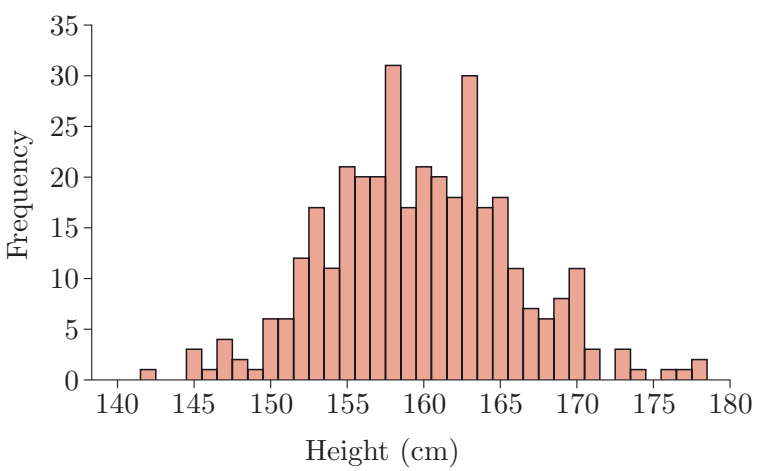


Figure 1 Heights of 351 elderly women

Example 2 *Chest measurements*

Scottish Highland regiments traditionally wear kilts as part of their dress uniform. It's believed that kilts were last widely worn in action at the Battle of Dunkirk in May 1940.

The data in Table 2 are the chest measurements (in inches to the nearest inch) of 5732 Scottish soldiers, measured in the early nineteenth century.

Table 2 Chest measurements of 5732 Scottish soldiers (inches)

Measurement	33	34	35	36	37	38	39	40
Frequency	3	19	81	189	409	753	1062	1082
Measurement	41	42	43	44	45	46	47	48
Frequency	935	646	313	168	50	18	3	1

(Source: Stigler, S.M. (1986) *The History of Statistics – The Measurement of Uncertainty Before 1900*, Cambridge, MA, Belknap Press of Harvard University Press, p. 208. The data are taken from the *Edinburgh Medical and Surgical Journal* (1817).)

The data are summarised in the histogram in Figure 2.

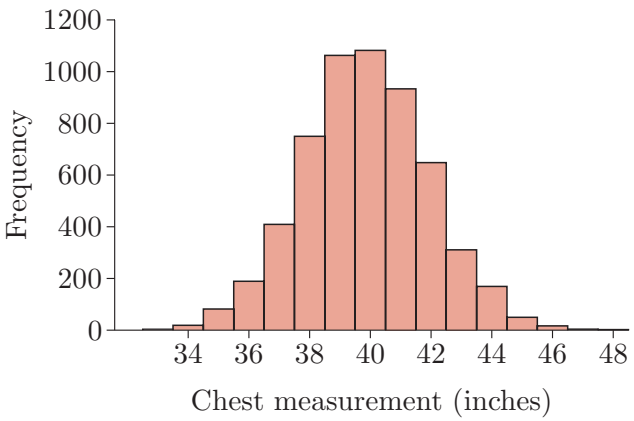


Figure 2 Chest measurements of 5732 Scottish soldiers

Activity 1 *Summarising common features*

The shapes of the histograms in Figures 1 and 2 are similar. Briefly describe their common shape.

Example 3 *Hepatitis*

For the datasets in Tables 3 and 4 (overleaf), the same random variable is being measured, but on two different sets of individuals. The variable is the logarithm of measures of ornithine carbonyltransferase (a liver enzyme) in patients suffering from one of two forms of hepatitis. Hepatitis is a term used to describe inflammation of the liver. The measurements in Table 3 are for 57 patients suffering from acute viral hepatitis (this includes five types of viral hepatitis, A to E); those in Table 4 are for 40 patients suffering from aggressive chronic hepatitis (now known as autoimmune hepatitis). The investigators compared the two groups of patients – specifically, they were interested in whether or not it was possible to distinguish between the patient groups on the basis of measurements of the enzyme.

Table 3 Enzyme measurements, acute viral hepatitis (log measure)

2.66	2.38	2.37	2.31	2.50	1.96	2.85	2.68	1.76	2.36	2.56	2.09
2.85	2.67	2.37	2.40	2.79	1.82	3.00	2.50	2.36	2.48	2.60	2.42
2.51	2.51	2.80	2.50	2.57	2.54	2.53	2.78	2.07	2.35	2.98	2.31
2.45	2.75	2.56	2.50	3.00	2.94	2.46	2.83	3.61	2.99	2.78	3.02
2.93	2.78	2.57	2.62	2.71	2.18	3.21	2.86	2.51			

Table 4 Enzyme measurements, aggressive chronic hepatitis (log measure)

3.01	2.99	2.60	2.47	3.04	1.92	2.17	2.33	2.07	2.30	2.56	2.11
3.32	2.21	1.71	2.60	2.79	2.71	2.64	2.52	2.21	2.58	2.40	2.45
3.18	2.84	2.84	2.31	2.71	2.47	2.72	3.71	2.73	3.69	3.40	2.77
2.28	2.84	2.80	3.02								

(Source: Albert, A. and Harris, E.K. (1987) *Multivariate Interpretation of Clinical Laboratory Data*, New York, Marcel Dekker Inc.)

Histograms for the two datasets are shown in Figure 3.

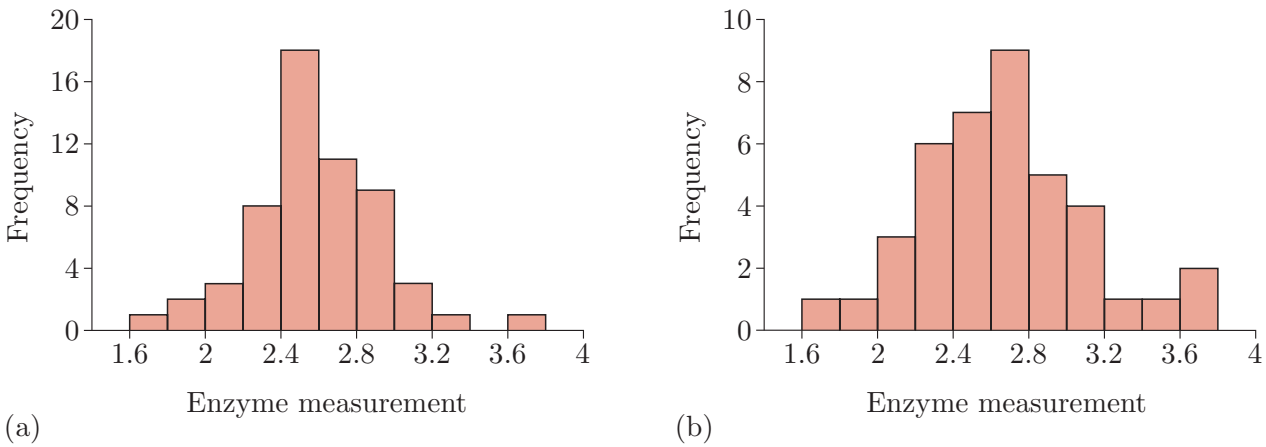


Figure 3 Enzyme measurements for (a) acute viral hepatitis and (b) aggressive chronic hepatitis

Since the measurements in Tables 3 and 4 are on the same variable, they can be compared directly. A comparative boxplot for the two groups of patients is shown in Figure 4.

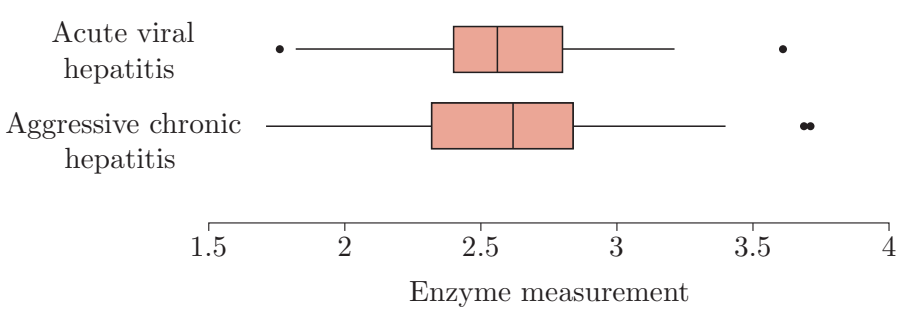


Figure 4 Comparison of the two sets of liver enzyme measurements on hepatitis patients

Activity 2 *Hepatitis*

What do the histograms and comparative boxplot in Figures 3 and 4 tell you about the two sets of measurements?

The common characteristics of the shapes of the histograms in Figures 1, 2 and 3 are shared with the p.d.f. of a normal distribution: normal distributions are symmetric about a single central mode, where the p.d.f. is at its maximum, and are sometimes described as 'bell-shaped'. The p.d.f. of a normal distribution, often called a 'normal curve', is shown in Figure 5.

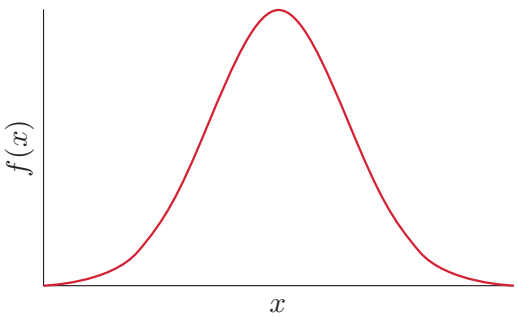
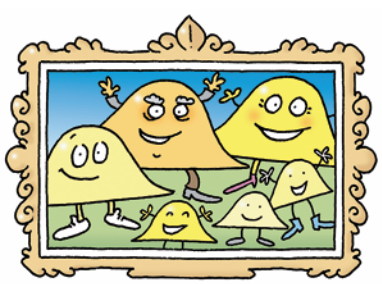


Figure 5 A normal curve

Note that there is more than one normal distribution. No single distribution could describe the data of Figure 1 (with mode at about 158 cm or so and values ranging from approximately 142 cm to 178 cm), and those of Figure 2 (with mode around 40 inches and values ranging from approximately 33 inches to 48 inches), and those of Figure 3 (with different modes and ranges of values again). In the real world, there are many instances of random variation following this kind of shape, although the mode and the spread of observed values alter from random variable to random variable. Different normal distributions are required to model the variation in different datasets.

The four panels of Figure 6 (overleaf) show normal curves superimposed on top of the unit-area histogram versions of Figures 1, 2 and 3 to show how the shapes of normal distributions compare with the shapes of the histograms for these datasets. Obviously the normal curves are not a perfect 'fit' to the shapes of the histograms, but they are close enough to suggest that normal distributions would be appropriate models for the data.

Normal distributions form a family of probability models. The normal family has two parameters: one specifying the location (the centre of the distribution) and one describing the spread (dispersion). The location parameter is denoted by μ, and the spread parameter is denoted by σ. The parameter μ can take any value; the parameter σ must be positive. The p.d.f. of a member of the normal family is symmetric about $x = \mu$ and is such that observations less than about $\mu - 3\sigma$ or more than about $\mu + 3\sigma$ are rather unlikely. A sketch of the normal p.d.f. showing how μ and σ relate to the curve is given in Figure 7 (overleaf). This is a rather important figure conveying lots of information about the normal distribution that you will use frequently in your study of statistics.



The Normal family?

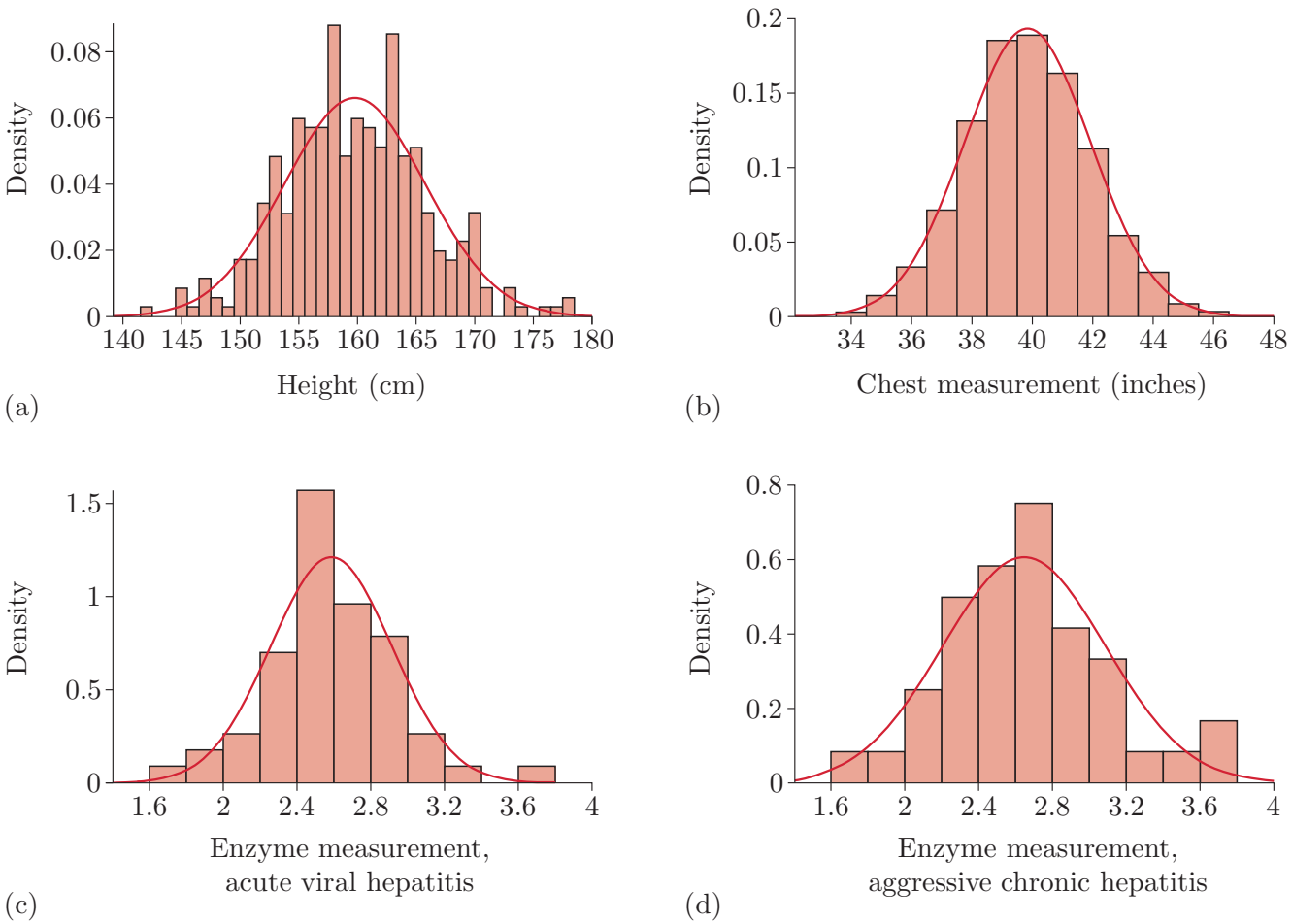


Figure 6 A histogram and normal curve for (a) the heights data given in Example 1, (b) the chest measurements data given in Example 2, (c) the acute viral hepatitis data given in Example 3 and (d) the aggressive chronic hepatitis data given in Example 3

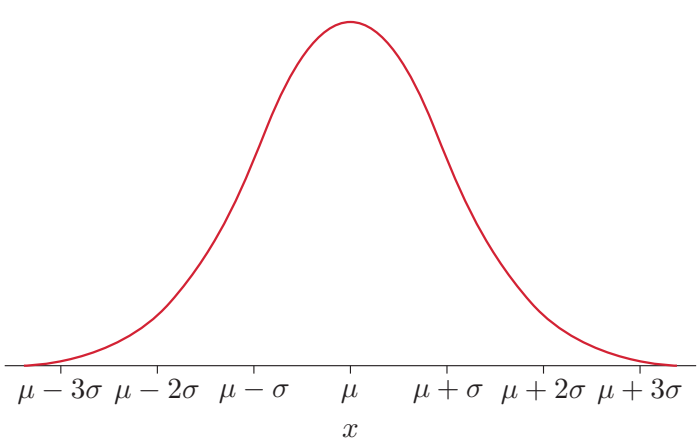


Figure 7 The p.d.f. of a normal distribution

Activity 3 *Sketching normal curves*

Using Figure 7 to guide you, sketch the curves for the following normal distributions.

(a) Normal distribution with $\mu = 10$ and $\sigma = 5$.

(b) Normal distribution with $\mu = -5$ and $\sigma = 2$.

(c) Normal distribution with $\mu = 0$ and $\sigma = 1$.

So how do the parameters μ and σ affect the shapes of normal distributions? From Figures 5, 6 and 7, the clear answer is that they do not! The *shape* of the normal p.d.f. is always the same: it is always unimodal and symmetric about $x = \mu$.

But μ and σ certainly do affect normal distributions by determining their location – that is, whereabouts they should be centred – and their spread or dispersion. You will explore how this works in the next activity.

Activity 4 *Exploring the effect of μ on the normal p.d.f.*

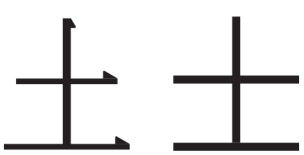
(a) Again use Figure 7 to guide you, paying particular attention to the p.d.f. becoming close to zero at $\mu \pm 3\sigma$, but this time draw all three sketches of normal curves on a common scale ranging from -30 to 30. Sketch the curves for the following normal distributions, each of which has the same value of σ.

(i) Normal distribution with $\mu = 0$ and $\sigma = 4$.

(ii) Normal distribution with $\mu = 10$ and $\sigma = 4$.

(iii) Normal distribution with $\mu = -10$ and $\sigma = 4$.

(b) What happens to the normal p.d.f. as μ increases but σ remains the same? What happens when μ decreases?



On the left, the Chinese character for 'scholar'; on the right, mathematical shorthand for 'plus or minus'

Activity 5 *Exploring the effect of σ on the normal p.d.f.*

(a) Again use Figure 7 to guide you, paying particular attention to the p.d.f. becoming close to zero at $\mu \pm 3\sigma$, but this time draw all three sketches of normal curves on a common scale ranging from -100 to 100. Sketch the curves for the following normal distributions, each of which has the same value of μ.

(i) Normal distribution with $\mu = 0$ and $\sigma = 10$.

(ii) Normal distribution with $\mu = 0$ and $\sigma = 25$.

(iii) Normal distribution with $\mu = 0$ and $\sigma = 5$.

(b) What happens to the normal p.d.f. as σ increases but μ remains the same? What happens when σ decreases?

The observations you have made in Activities 4 and 5 are quite general. Whatever the value of σ, increasing μ has the effect of moving the normal p.d.f. to the right, and decreasing μ has the effect of moving the normal p.d.f. to the left. Also, whatever the value of μ, increasing σ has the effect of making the normal p.d.f. look wider and flatter, and decreasing σ has the effect of making the normal p.d.f. look narrower and taller. When μ and σ both vary, horizontal movement and widening/narrowing effects occur simultaneously.

Hence justifying the notation!

Now, it can be shown that μ is the mean of the normal distribution and σ is its standard deviation (so that σ^2 is its variance). These properties of the normal distribution will not be proved in this module, but are sufficiently important to stress in the following box.

The mean, standard deviation and variance of a normal distribution

If the random variable X has a normal distribution with parameters μ and σ, then the mean, standard deviation and variance of X are given by

$$E(X) = \mu, \quad S(X) = \sigma \quad \text{and} \quad V(X) = \sigma^2.$$

The normal distribution itself is defined in the box below.

The normal distribution

The continuous random variable X is normally distributed with mean μ and standard deviation σ (and hence variance σ^2) if the probability density function of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty.$$

The normal distribution is often called the Gaussian distribution.

This is written $X \sim N(\mu, \sigma^2)$.

Observe that the notation $N(\mu, \sigma^2)$ gives the mean μ and the *variance* σ^2 of the normal distribution, not the mean and standard deviation: for instance, if we write $X \sim N(3, 4)$, then 4 is the *variance* σ^2 of X, and the standard deviation σ is $\sqrt{4} = 2$.

Notice also that for a normal distribution all values are possible, both negative and positive, whereas in the real world there are very few quantities for which observations may include all negative and positive numbers. However, for any normal distribution, values more than about three standard deviations from the mean are unlikely. So extreme values may be regarded as occurring with negligible probability; hence a normal model may be suitable even when the quantity being modelled takes only a limited range of values. It follows that you should not make a statement such as 'the variation in the chest measurements of Scottish soldiers is normally distributed with mean 40 inches and standard deviation about

2 inches' as this implies that observations of any size are possible, including negative observations. Rather, you should say 'the variation in the chest measurements of Scottish soldiers *may be adequately, or reasonably, modelled by* a normal distribution with mean 40 inches and standard deviation about 2 inches'.

We will return to the question of estimating the parameters of a normal distribution, and hence deciding which normal distribution is most appropriate as a model for the variation in chest measurements and other datasets, later in the unit.

Exercise on Section 1

Exercise 1 *Comparing normal p.d.f.s*

In each part of this exercise, state how the p.d.f. associated with Y differs from the p.d.f. associated with X.

(a) $X \sim N(1, 1)$, $Y \sim N(0.5, 0.5)$

(b) $X \sim N(-1, 3)$, $Y \sim N(-1, 1)$

(c) $X \sim N(0, 10)$, $Y \sim N(-1, 100)$

(d) $X \sim N(-5, 1)$, $Y \sim N(0, 1)$

2 Calculating probabilities

For a normal distribution, as for other continuous probability distributions, probabilities are calculated by finding areas under the graph of the p.d.f. As previously observed, using the c.d.f. is the easiest way to do this. Example 4 and Activity 6 illustrate the use of the c.d.f. in situations where a normal model is proposed.

Example 4 *Chest measurements of Scottish soldiers*

Let us suppose that a normal distribution with parameters $\mu = 40$ and $\sigma = 2$ (so that $\sigma^2 = 4$) is an adequate model for the chest measurements (in inches) of nineteenth-century Scottish soldiers represented in Figure 2. This model may be used to estimate, for the population of Scottish soldiers from which the sample was drawn, the proportion of the population with chest measurements in any given range. For example, we can estimate the proportion of the population of Scottish soldiers whose chests measured between 37 and 42 inches inclusive.

You have seen that a helpful first step in any calculation involving a continuous random variable is to draw a rough sketch showing the area representing the proportion or probability required. In this case, if X is a random variable representing the chest measurement of a Scottish soldier,

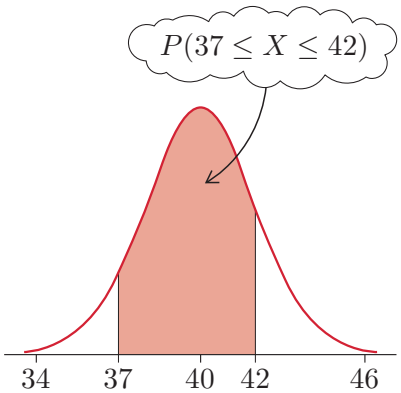


Figure 8 The p.d.f. of $X \sim N(40, 4)$ showing $P(37 \leq X \leq 42)$

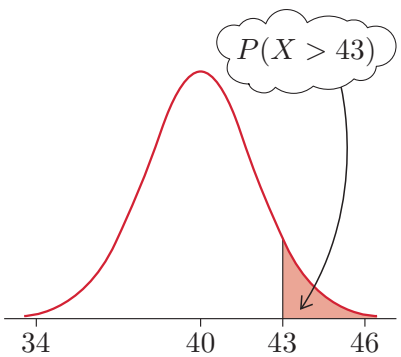


Figure 9 The p.d.f. of $X \sim N(40, 4)$ showing $P(X > 43)$

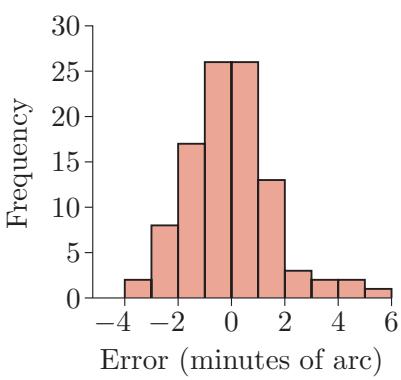


Figure 10 Errors in angular measurements

then $X \sim N(40, 4)$ and the probability required is $P(37 \leq X \leq 42)$. This probability is given by the shaded area in Figure 8.

This area is equal to the total area to the left of 42 minus the area to the left of 37; that is,

$$P(37 \leq X \leq 42) = F(42) - F(37),$$

where F is the c.d.f. of X.

The shaded area in Figure 9 represents the proportion of Scottish soldiers who, according to the model, had chest measurements greater than 43 inches. Since the total area under a p.d.f. is equal to 1, this area is equal to 1 minus the area to the left of 43, that is,

$$P(X > 43) = 1 - F(43).$$

So questions about the proportion of the population with chest measurements in any given range may be answered by finding values of the c.d.f. of the normal model.

Activity 6 *Errors in angular measurements*

During the mapping of the state of Massachusetts in the USA in the nineteenth century, 100 readings were taken on the error involved when measuring angles. The data, which are from the United States Survey Report (1854), are represented in Figure 10. The error was measured in minutes of arc. (A minute is 1/60 of a degree.)

A normal distribution with mean 0 and variance 2.75 is a reasonable model for the angular measurement errors. If X is a random variable representing the error in an angular measurement, then for each of the following probabilities, draw a sketch showing the area which represents the probability, and express the probability in terms of the c.d.f. of X.

(a) The probability that an error is positive and greater than 0.5 minutes of arc.

(b) The probability that an error is positive but less than 2 minutes of arc.

(c) The probability that the size of an error, which may be positive or negative, is less than 1 minute of arc.

(d) The probability that the size of an error, positive or negative, is greater than 1.5 minutes of arc.

Example 5 *More on chest measurements*

In Example 4, the c.d.f. of a normal distribution was used to write down expressions for the proportion of nineteenth-century Scottish soldiers whose chest measurements were, according to the model, between 37 and 42 inches, and for the proportion whose chest measurements were greater than 43 inches.

Now suppose that you want to know the value of x such that only 2.5% of Scottish soldiers' chest measurements were less than x inches, and the value of y such that only 5% of Scottish soldiers' chest measurements exceeded y inches. These are the sorts of questions that can be answered by finding appropriate quantiles of the normal model, $X \sim N(40, 4)$.

The chest measurement x below which only 2.5% of measurements lie is, according to the model, the solution of

$$P(X \leq x) = 0.025.$$

That is, $F(x) = 0.025$, so $x = q_{0.025}$, the 0.025-quantile of $N(40, 4)$. This is illustrated in Figure 11.

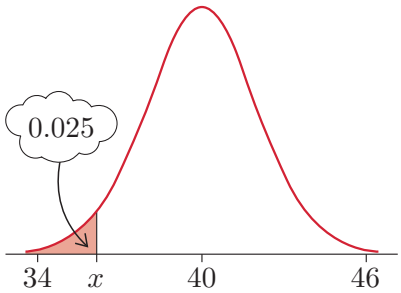


Figure 11 The 0.025-quantile of $N(40, 4)$

Similarly, according to the model, 5% of chest measurements were greater than y, where

$$P(X > y) = 0.05.$$

So

$$F(y) = P(X \leq y) = 0.95,$$

and hence the required value y is $q_{0.95}$, the 0.95-quantile of $N(40, 4)$. This is illustrated in Figure 12.

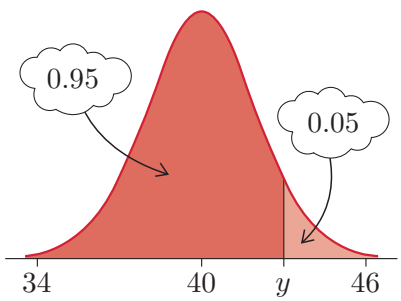


Figure 12 The 0.95-quantile of $N(40, 4)$

Activity 7 *More on angular errors*

In Activity 6, a normal distribution with mean 0 and variance 2.75 was proposed as a model for the errors in angular measurements represented in Figure 10.

(a) Suppose that 10% of errors are positive and greater than x. Draw a sketch showing this information. Express x as a quantile of $N(0, 2.75)$.

(b) Suppose that 95% of errors lie between $-b$ and b (where b is positive). Draw a sketch showing this information. Express b as a quantile of $N(0, 2.75)$.

(c) Suppose that 99% of errors lie between $-c$ and c (where c is positive). Draw a sketch showing this information. Express c as a quantile of $N(0, 2.75)$.

Examples 4 and 5 and Activities 6 and 7 illustrate how the c.d.f. may be used to solve problems assuming a normal model. Given a formula for the c.d.f., we could go on to calculate probabilities and find quantiles for the given normal models. Unfortunately, there is no useful explicit formula for the c.d.f. of a normal distribution. So, in practice, values of the c.d.f. and quantiles of a normal distribution are found either using printed tables or using a statistical software package. You will learn how to use printed tables in Section 4. The use of Minitab to find values of the c.d.f. and quantiles for normal distributions is described in Chapter 1 of Computer Book B.

Refer to Chapter 1 of Computer Book B for the work on calculating probabilities and quantiles of normal distributions using Minitab.

The results you obtained in Chapter 1 of Computer Book B are all special cases of the following general result.

A general result for normal distributions

If a normal distribution is used to model the variation in a population, then according to the model, the proportion of the population within k standard deviations of the mean is the same, whatever the values of the mean μ and the standard deviation σ.

Equivalently, if $X \sim N(\mu, \sigma^2)$, then the probability $P(\mu - k\sigma < X < \mu + k\sigma)$ depends only on the value of k, and not on μ and σ.

The importance of this result will become evident in Section 4: it enables calculations for any normal distribution to be done using printed tables for the normal distribution with mean 0 and standard deviation 1.

Exercise on Section 2

Exercise 2 *IQ test*

When asked what his IQ is, Stephen Hawking said: 'I have no idea. People who boast about their I.Q. are losers.' (*New York Times*, 12 December 2004)

One particular type of IQ test is designed so that scores, X, on the test are normally distributed with mean 100 and standard deviation 15.

(a) Sketch the normal curve for the IQ scores.

(b) Draw a sketch showing the area which represents the probability that a score on the IQ test is between 110 and 130. Express this area in terms of the c.d.f. of X.

(c) Suppose that 5% of scores are greater than x. Express x as a quantile of $N(100, 15^2)$.

3 Linear functions of normal random variables

In Unit 4, we looked at the means and variances of both a linear function of a single random variable and a sum of independent random variables. Often, the random variable or variables involved can be assumed to follow a normal distribution. In this case, we can do more than just obtain means and variances: we can obtain the entire distribution of the linear function and of the sum. This is the topic of this section.

Examples 6 and 7 exemplify two of the many situations in which the distributions of a linear function of a normal random variable or of a sum of independent normal random variables are required.

Example 6 *Chest measurements in centimetres*

In Example 12 of Unit 4, we considered the following situation. We let X be a random variable representing a chest measurement made in inches, and let Y be a random variable representing the same chest measurement made in centimetres. We noted that X and Y are (approximately) related by the linear function

$$Y = 2.54X.$$

We also showed, in Examples 15 and 16 of Unit 4, that if X followed a distribution with mean $E(X) = 40$ and standard deviation $S(X) = 2$ (inches), then

$$E(Y) = 2.54\,E(X) = 101.6 \quad \text{and} \quad V(Y) = 2.54^2\,V(X) \simeq 25.81$$

(centimetres).

In Example 4 of this unit, we claimed that a reasonable model for the chest measurements in inches of Scottish soldiers in the early nineteenth century is given by a normal distribution with mean $\mu = 40$ and standard deviation $\sigma = 2$ inches. Given the extra information (compared with Unit 4) that X is normally distributed, what can we say about the whole distribution of chest measurements in centimetres? In particular, if a normal distribution is a good model for X, might a normal distribution also be a good model for Y?

Despite the UK government agreeing to support metrification as long ago as 1965, imperial measurements are still widely used in the UK today

---

Example 7 *Bags of sugar*

Sugar is sold in bags in a variety of sizes. Consider those labelled as containing 2 kg. There is some variability in the contents of bags, so the amount of sugar (in grams) in a bag is a random variable X. Suppose that a reasonable model for X is given by a normal distribution: $X \sim N(2003, 10)$.

A cook who requires 6 kg of sugar for making marmalade buys three bags. What is the probability that he has less than 6 kg of sugar?

To answer this question, we need to know the distribution of the total contents of three randomly selected bags of sugar. That is, if X_1, X_2 and X_3 represent the contents in grams of three randomly selected bags, then we need to know the distribution of the random variable $S = X_1 + X_2 + X_3$. Is S also normally distributed?

3.1 The distribution of a linear function of a normal random variable

Recall from Subsection 5.1 of Unit 4 that for any random variable X and constants a and b,

$$E(aX + b) = a\,E(X) + b \tag{1}$$

and

$$V(aX + b) = a^2\,V(X). \tag{2}$$

Activity 8 *Sections of a chemical reactor revisited*

One of the world's largest chemical industry sites, in Ludwigshaven, Germany

In Exercise 11 of Unit 4, you considered the variation in temperature across sections of a chemical reactor. Here, you will consider a 'reversed' version of the question of interest and an extension thereof. Suppose now that the water temperature measured in °C (degrees Celsius) is represented by a random variable X. We will assume that variation in the section temperature may be adequately modelled by a normal distribution with mean 452 and standard deviation 22 (°C).

Suppose that Y is a random variable representing the water temperature measured in °F (degrees Fahrenheit). Then

$$Y = \frac{9}{5}X + 32.$$

What are the mean and variance of the section temperature measured in °F? Can you conjecture what the distribution of Y might be?

In Activity 8, the temperature in °C was modelled by a normal distribution. You found the mean and standard deviation of Y, the temperature in °F, but is a normal distribution also an adequate model for the new random variable? It turns out that Y is also normally distributed. Altering the scale of measurement does not change the essential characteristics of the temperature variation: the probability density function reaches a peak at the mean temperature and tails off symmetrically either side of the mean. The fact that Y is also normally distributed follows from a general result which will not be proved here: if a random variable is normally distributed, then any linear function of the random variable is also normally distributed.

The distribution of a linear function of a normal random variable

If the random variable X is normally distributed with mean μ and variance σ^2, and a and b are constants, then the random variable $Y = aX + b$ is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$:

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2). \tag{3}$$

Example 8 *Chest measurements revisited*

Distributional Result (3) may be used to answer the question posed in Example 6 at the beginning of this section. A normal distribution with mean $\mu = 40$ and standard deviation $\sigma = 2$ was proposed for X so that $X \sim N(40, 4)$, so since $Y = 2.54X$, using Distributional Result (3) gives

$$Y \sim N(101.6, 25.81).$$

Activity 9 *Temperature in Dallas*

Suppose that the temperature measured in °F at 10 a.m. on a July day in Dallas, Texas, USA, can be modelled by a normal distribution with mean 84 and standard deviation 4. Which distribution is an adequate model for the temperature at 10 a.m. on a July day in Dallas as measured in °C? Hint: you will find the equation linking X and Y in Activity 8 useful.

According to Wikipedia, the highest recorded temperature in the USA was 134 °F/57 °C on 10 July 1913, in Death Valley, California

3.2 The distribution of a sum of independent normal random variables

You have seen in Unit 4 that the mean of a sum of random variables is equal to the sum of the means of the random variables,

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n), \tag{4}$$

and that if the random variables are independent, then the variance of their sum is equal to the sum of their variances,

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n). \tag{5}$$

For normal random variables we have the further result that a sum of independent normal random variables has a normal distribution. The result is stated below.

The distribution of a sum of independent normal random variables

If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then their sum $Y = X_1 + X_2 + \cdots + X_n$ has a normal distribution with mean $\mu_1 + \mu_2 + \cdots + \mu_n$ and variance $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$. That is,

$$Y \sim N(\mu_1 + \mu_2 + \cdots + \mu_n,\ \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2). \tag{6}$$

The fact that the sum is normally distributed is stated without proof.

Example 9 *Bags of sugar*

On 5 February 2015, the *Washington Post* reported that the USA has the highest average sugar consumption per person in the world, followed by Germany and then the Netherlands

In Example 7, a situation was described in which a cook who requires 6 kg of sugar buys three bags, each of which is labelled as containing 2 kg. A normal model was assumed for X, the amount of sugar (in grams) in a bag: $X \sim N(2003, 10)$. What is the probability that the cook has less than 6 kg of sugar?

We are now in a position to write down the distribution of $S = X_1 + X_2 + X_3$, where X_1, X_2, X_3 are the contents of three randomly selected bags. Since the bags are selected randomly, we can assume that X_1, X_2 and X_3 are independent random variables, each having a normal distribution with mean 2003 and variance 10. So, using Distributional Result (6), the distribution of S, the total contents of three bags, is normal with mean

$$E(S) = E(X_1) + E(X_2) + E(X_3) = 2003 + 2003 + 2003 = 6009;$$

and the variance of S is

$$V(S) = V(X_1) + V(X_2) + V(X_3) = 10 + 10 + 10 = 30.$$

Thus the probability that the cook has less than 6 kg of sugar is given by

$$P(S < 6000) = F(6000),$$

where F is the c.d.f. of $S \sim N(6009, 30)$. The value of this probability may be found either using a computer or using printed tables. This problem is revisited in the next section.

Activity 10 *Travelling to work*

According to the *Daily Express* (29 July 2014), Thailand has the longest commuting times in the world, with the average working person spending 2 hours every day travelling to and from work

Jack walks the first part of his route to work, and then catches a bus. The time T_1 (in minutes) that he takes to walk to the bus stop may be reasonably modelled by a normal distribution: $T_1 \sim N(15, 0.5)$. The time T_2 (in minutes) after he reaches the bus stop until he alights from a bus at his workplace may also be reasonably modelled by a normal distribution: $T_2 \sim N(20, 9)$.

(a) Assuming that the walking time T_1 and the bus time T_2 are independent, write down the distribution of the total time it takes Jack to get to work.

(b) Give an expression for the probability that Jack will be late for work if he leaves home 40 minutes before he is due at work.

3.3 The distribution of a difference of independent normal random variables

To end this section, we consider briefly the difference $X - Y$ of two independent normal random variables X and Y.

The key to using results you already know is to write

$$X - Y = X + (-Y).$$

In the next activity, you will use this relationship to obtain the mean and variance of $X - Y$.

Activity 11 *$E(X - Y)$ and $V(X - Y)$*

(a) Use Equations (1) and (2) to write $E(-Y)$ and $V(-Y)$ in terms of $E(Y)$ and $V(Y)$.

(b) Use Equations (4) and (5) to write $E(X - Y)$ and $V(X - Y)$ in terms of $E(X)$, $E(Y)$, $V(X)$ and $V(Y)$.

If X and Y are also normally distributed, then by Distributional Result (3), $-Y$ is normally distributed; so, applying Distributional Result (6), $X - Y = X + (-Y)$ is also normally distributed. These results may be summarised as follows.

The distribution of a difference of independent normal random variables

If X and Y are independent normal random variables, then the distribution of $X - Y$ is also normal, and

$$E(X - Y) = E(X) - E(Y), \tag{7}$$
$$V(X - Y) = V(X) + V(Y). \tag{8}$$

Note that the result that $E(X - Y) = E(X) - E(Y)$ holds for any random variables, whether or not they are independent or normal. However, the independence condition is essential both for Equation (8) for the variance of $X - Y$ and for $X - Y$ to be normally distributed. Bear in mind that the variance of the difference between independent normal random variables is the *sum* of their variances (not their difference).

Activity 12 *Height differences*

In a particular population, the heights in centimetres of adult men are modelled by $N(172, 19)$, and the heights of adult women are modelled by $N(163, 13)$. A man and a woman are selected at random from the population.

(a) What is the distribution of the difference between the man's height and the woman's height?

(b) Write down an expression for the probability that the man is more than 8 cm taller than the woman.

(c) Write down an expression for the probability that the woman is taller than the man.

Exercises on Section 3

Exercise 3 *A linear function of a random variable*

The random variable X is normally distributed with mean 10 and variance 5. What is the distribution of the random variable Y which is given by $Y = 3X + 2$?

Exercise 4 *Sum of independent normal random variables*

Three independent random variables X_1, X_2 and X_3 are all normally distributed so that $X_1 \sim N(-10, 2)$, $X_2 \sim N(15, 12)$ and $X_3 \sim N(-2, 5)$. What is the distribution of $S = X_1 + X_2 + X_3$?

Exercise 5 *A difference*

Suppose that X and Y are independent normal random variables so that $X \sim N(-5, 4)$ and $Y \sim N(10, 5)$. What is the distribution of $X - Y$?

4 Calculations using tables

Even though statistics software packages are available to do calculations for normal distributions, printed tables are still used when it is not convenient, or not worthwhile, to use a computer. Subsection 4.1 considers the normal distribution with mean 0 and standard deviation 1; this distribution is called the *standard normal distribution*. The use of printed tables to find probabilities and quantiles for the standard normal distribution is explained in Subsections 4.2 and 4.3, respectively. In Subsection 4.4, a procedure known as *standardisation* is described. You will see how, using this procedure, any problem about a normal distribution may be expressed as a problem about a normal distribution with mean 0 and standard deviation 1, so that tables for the standard normal distribution may be used to find probabilities and quantiles for *any* normal distribution.

4.1 The standard normal distribution

At the end of Section 2, it was stated that if a normal distribution is used to model the variation in a population, then according to the model, the proportion of the population within k standard deviations of the mean is the same, whatever the values of the mean μ and the standard deviation σ.

This result means that the proportion of the whole area (which is 1) that is shaded in Figure 13(a) depends on the value of k but not on the values of μ and σ. In particular, for any values of μ and σ, this area is equal to the area between $-k$ and k under the curve of a normal distribution with mean 0 and standard deviation 1. This is illustrated in Figure 13(b).

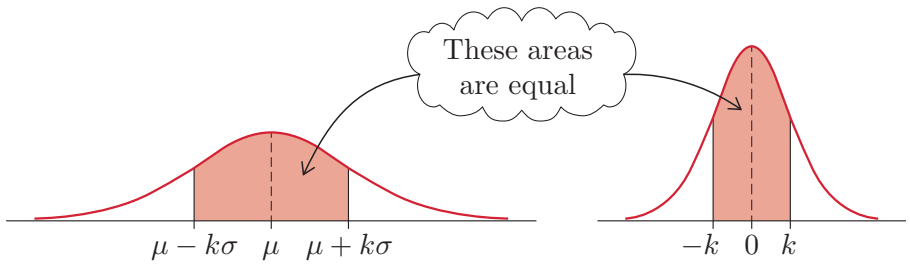


Figure 13 The proportion within k standard deviations of the mean for (a) $N(\mu, \sigma^2)$ and (b) $N(0, 1)$

This actually means that any problem involving a normal distribution may be expressed as a problem about a normal distribution with mean 0 and standard deviation 1. (A procedure for doing this is described in Subsection 4.4.) The normal distribution with mean 0 and standard deviation 1, which therefore has a particularly important role in statistics, is called the **standard normal distribution**. It is defined in the box below.

The standard normal distribution

The standard normal distribution is the normal distribution with mean 0 and standard deviation 1. The letter Z is usually used to denote the standard normal random variable: $Z \sim N(0, 1)$.

The p.d.f. of Z is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty. \tag{9}$$

The c.d.f. of Z is given by

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx. \tag{10}$$

ϕ is the Greek lower-case letter phi, and Φ is the Greek upper-case letter phi. Both are pronounced 'fye'.

The letter ϕ is conventionally used for the p.d.f. of Z, and the letter Φ for the c.d.f. of Z.

The graph of the p.d.f. of Z is shown in Figure 14(a) (overleaf). The p.d.f. is positive for any value of z but, as you can see, observations much less than -3 or greater than $+3$ are unlikely. The c.d.f. is represented on the plot of the p.d.f. in Figure 14(b): the c.d.f. is the area to the left of z under the standard normal curve, which is shaded in the diagram; it is equal to $\Phi(z)$.

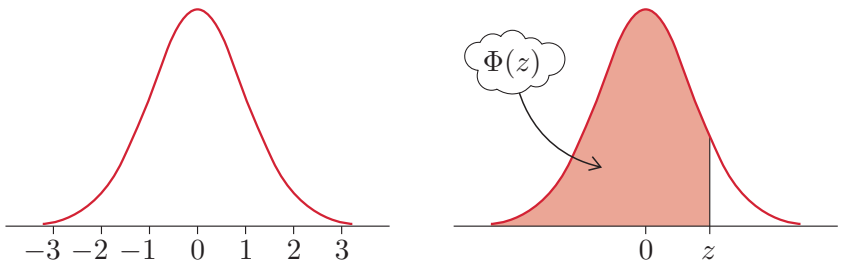
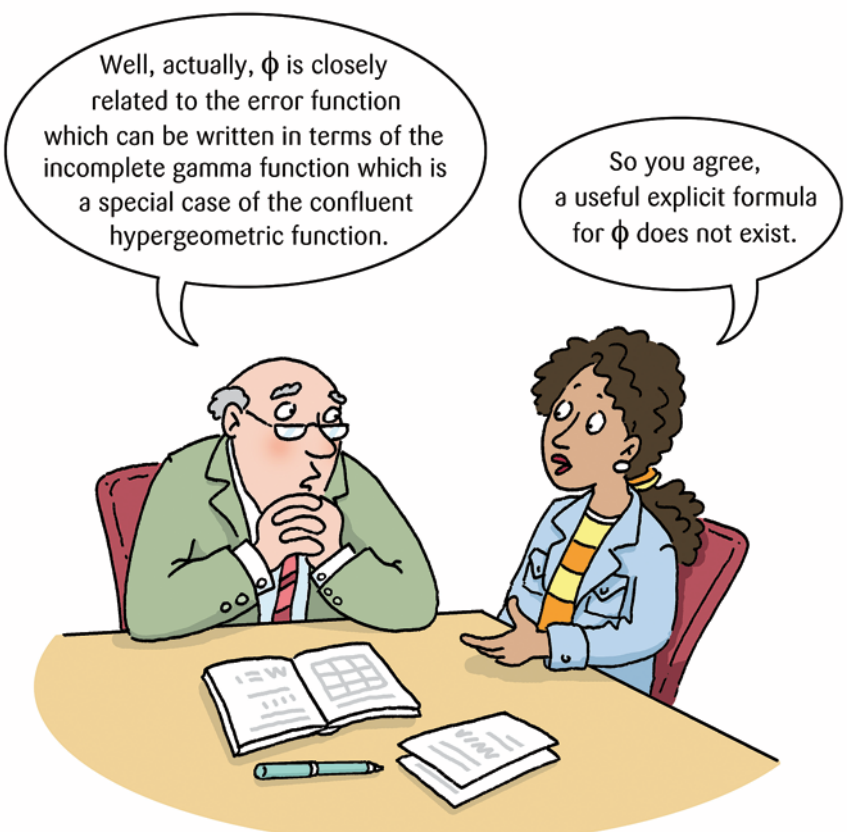


Figure 14 (a) The p.d.f. and (b) its relationship with the c.d.f. of $Z \sim N(0, 1)$

As already observed, a useful explicit formula for the c.d.f., $\Phi(z)$, of a normal distribution does not exist; all we have, in Equation (10), is its expression as an integral. So, instead of using a formula to find values of $\Phi(z)$, printed tables or a computer are used to obtain values. In Chapter 1 of Computer Book B, you learned how to use Minitab to find values of the c.d.f. and to find quantiles for any normal distribution. The use of tables to find values of the c.d.f. of, and hence probabilities associated with, the standard normal distribution is explained in Subsection 4.2; the use of tables to find quantiles of the standard normal distribution is explained in Subsection 4.3.



4.2 Using printed tables to obtain probabilities

Tables for the standard normal distribution take many forms. For example, some contain values of $P(0 \leq Z \leq z)$; other tables give values of $P(Z \geq z)$ or even $P(-z \leq Z \leq z)$. There are so many variations on questions that might be asked that no particular table is more convenient than any other. Table 5 (overleaf), which is reproduced in the M248 Handbook, gives values of the c.d.f. of the standard normal distribution: for values of z from 0 to 4.09 (in steps of 0.01), it gives values of

$$\Phi(z) = P(Z \leq z).$$

The printed values are accurate to four decimal places.

A different kind of printed table

You may find it more convenient to refer to the table in the Handbook than to Table 5 as you work through the next few examples and activities.

Example 10 *Using tables for the c.d.f. of Z*

The probability $P(Z \leq 1.58) = \Phi(1.58)$ is given by the entry in Table 5 in the row labelled 1.5 in the column headed 8; this is 0.9429.

Similarly, the probability $P(Z \leq 0.73) = \Phi(0.73)$ is given by the entry in the row labelled 0.7 in the column headed 3; this is 0.7673.

These probabilities are illustrated on sketches of the standard normal p.d.f. in Figure 15.

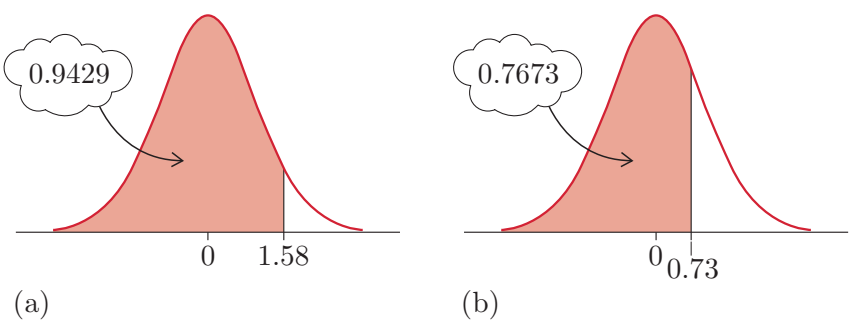


Figure 15 (a) $P(Z \leq 1.58)$, (b) $P(Z \leq 0.73)$

Activity 13 *Using the table*

Use Table 5 to find each of the following probabilities.

(a) $P(Z \leq 1.00)$

(b) $P(Z \leq 1.96)$

(c) $P(Z \leq 2.25)$

In each case, illustrate the probability on a sketch of the standard normal p.d.f.

Table 5 Probabilities for the standard normal distribution $\Phi(z) = P(Z \leq z)$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Of course, required probabilities will not always be of the form $P(Z \leq z)$, where $z \geq 0$. However, by making use of the following two properties, we can calculate any probability for the standard normal distribution that we might require.

- The standard normal distribution is symmetric about 0.
- The total area under the p.d.f. of the standard normal distribution is 1.

The calculation of probabilities such as $P(Z \geq 1.50)$ or $P(0 \leq Z \leq 1.83)$ or $P(-1.33 \leq Z \leq 2.50)$ are illustrated in the next few examples.

Example 11 *More on using the table*

Since Φ is the c.d.f. of the standard normal distribution, $\Phi(z) = P(Z \leq z)$ is equal to the area to the left of z under the standard normal p.d.f. So when calculating probabilities, it usually helps to draw a rough sketch showing the area under the standard normal p.d.f. that is required. The shaded area in Figure 16 represents $P(Z \geq 1.50)$.

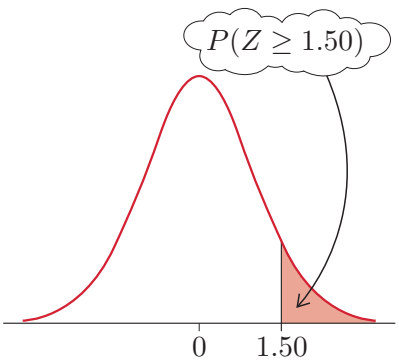


Figure 16 $P(Z \geq 1.50)$

From the table,

$$\Phi(1.50) = P(Z \leq 1.50) = 0.9332.$$

Since the total area under the curve is 1, it follows by subtraction from 1 that

$$P(Z \geq 1.50) = 1 - \Phi(1.50) = 1 - 0.9332 = 0.0668.$$

This is illustrated in Figure 17(a).

Since Z is continuous,

$$P(Z < 1.50) = P(Z \leq 1.50) = \Phi(1.50).$$

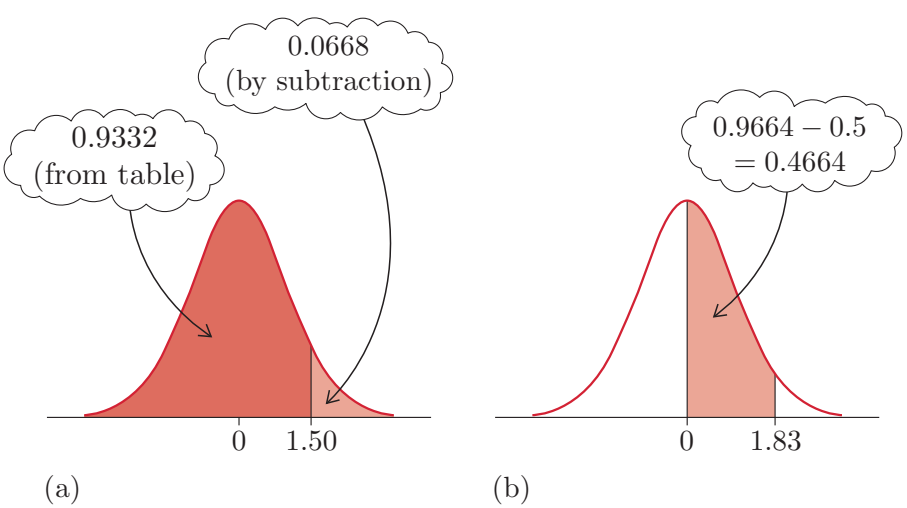


Figure 17 Finding (a) $P(Z \geq 1.50)$, (b) $P(0 \leq Z \leq 1.83)$

The sketch in Figure 17(b) shows the area which represents the probability $P(0 \leq Z \leq 1.83)$. This area may be found by first noting that it is equal to the area to the left of 1.83 minus the area to the left of 0. Then

$$P(0 \leq Z \leq 1.83) = \Phi(1.83) - \Phi(0) = 0.9664 - 0.5000 = 0.4664.$$

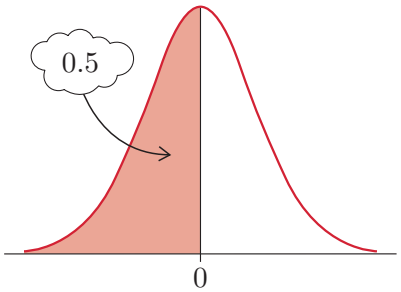


Figure 18 $P(Z \leq 0)$

Note that

$$\Phi(0) = P(Z \leq 0) = 0.5.$$

This is true because of the symmetry of the standard normal p.d.f.; see Figure 18. (It is also given as the first entry in Table 5.) It follows that $P(Z \geq 0) = 1 - \Phi(0) = 0.5$ too.

Activity 14 *More practice at using the table*

For each of the following probabilities, draw a rough sketch of the p.d.f. of the standard normal distribution and mark on your sketch the area which represents the probability. Then use Table 5 to find the value of the probability.

(a) $P(Z \geq 0.82)$

(b) $P(0.50 \leq Z \leq 1.50)$

Snowflakes also display considerable symmetry

Table 5 gives values of $\Phi(z)$ only for values of z between 0 and 4.09. Since $\Phi(4) = 1.0000$ to four decimal places, it follows that $\Phi(z) = 1.0000$ to four decimal places for any value of z greater than 4. But how can the table be used to find values of $\Phi(z)$ when z is negative? The answer is to use the symmetry of the standard normal curve.

Example 12 *Using the symmetry of the standard normal curve*

The area under the p.d.f. of the standard normal curve to the left of -1 is $\Phi(-1)$. Since the curve is symmetric about $z = 0$, this area is equal to the area under the p.d.f. to the right of $+1$. So

$$P(Z \leq -1) = P(Z \geq 1).$$

This is illustrated in Figure 19(a).

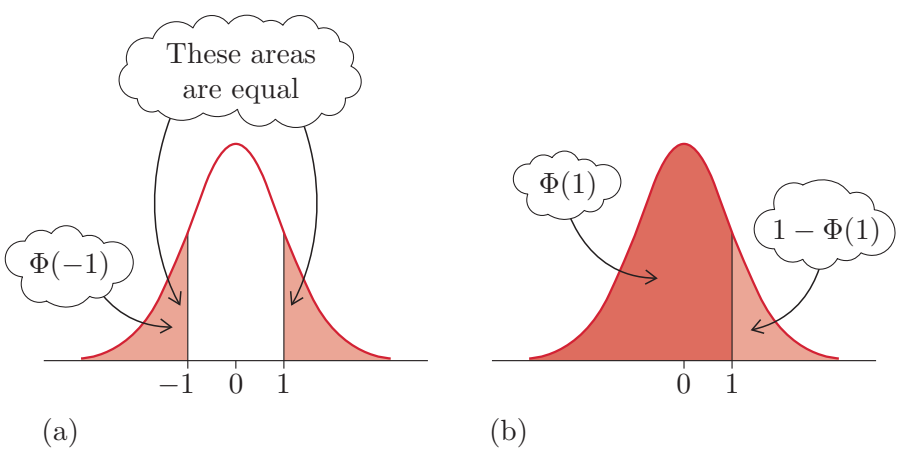


Figure 19 Using symmetry to calculate probabilities

From Figure 19(b), you can see that the shaded area on the right is equal to $1 - \Phi(1)$. So

$$\Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587.$$

A generalisation of this result is given in the following box and illustrated in Figure 20.

Using the symmetry of the standard normal distribution

For any $z > 0$,

$$P(Z \leq -z) = P(Z \geq z),$$

so

$$\Phi(-z) = 1 - \Phi(z).$$

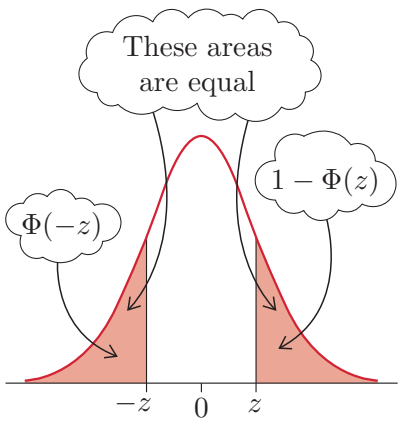


Figure 20 Using symmetry

Example 13 *Calculating probabilities*

The area representing the probability $P(-1.33 \leq Z \leq 2.50)$ is shown in Figure 21. This area is equal to

$$\begin{aligned}\Phi(2.50) - \Phi(-1.33) &= \Phi(2.50) - (1 - \Phi(1.33)) \\ &= 0.9938 - (1 - 0.9082) \\ &= 0.9938 - 0.0918 = 0.9020.\end{aligned}$$

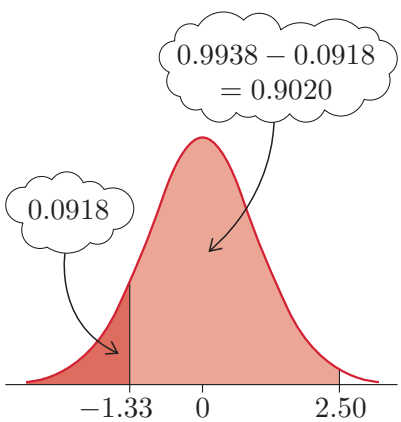


Figure 21
$P(-1.33 \leq Z \leq 2.50)$

Example 14 *More than three standard deviations from the mean*

In Chapter 1 of Computer Book B, you found that for several normal distributions, the proportion of the population being modelled that are more than three standard deviations from the mean is equal to 0.0027. Table 5 can be used to check that this is also the case for a standard normal distribution:

$$\begin{aligned}P(|Z| > 3) &= P(Z < -3) + P(Z > 3) \\ &= 2 \times P(Z < -3) = 2 \times \Phi(-3) = 2 \times (1 - \Phi(3)) \\ &= 2 \times (1 - 0.9987) = 0.0026.\end{aligned}$$

The slight discrepancy in values is due to rounding in the calculations.

$P(|Z| > 3)$ also equals $2 \times P(Z > 3)$.

Activity 15 *Using symmetry*

For each of the following probabilities, draw a rough sketch of the p.d.f. of the standard normal distribution and mark on your sketch the area that represents the probability. Then use Table 5 to find the value of the probability.

(a) $P(Z < -0.66)$

(b) $P(-2.67 < Z < 0.33)$

(c) $P(|Z| \leq 1.62) = P(-1.62 \leq Z \leq 1.62)$

(d) $P(|Z| \geq 2.45)$

(e) $P(-2.49 \leq Z < -0.65)$

Screencast 6.1 reviews using the standard normal distribution table to find probabilities of normal random variables.

Screencast 6.1 Using the standard normal table to find probabilities

4.3 Using printed tables to obtain quantiles

Let us now turn our attention to finding quantiles of the standard normal distribution.

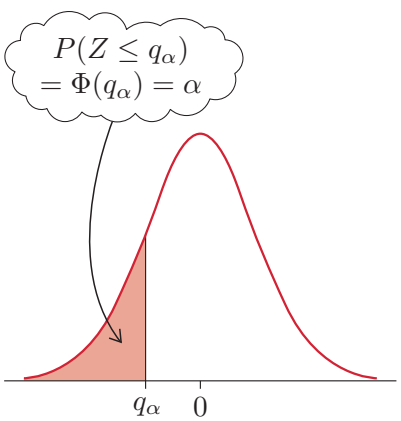


Figure 22 α-quantile, q_α, for $N(0, 1)$

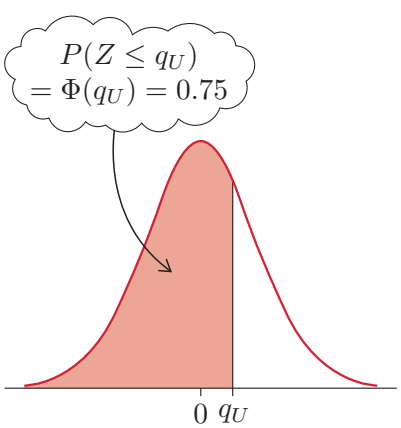


Figure 23 q_U for $N(0, 1)$

Example 15 *Finding quantiles of Z*

Recall from Subsection 4.1 of Unit 5 that for $0 < \alpha < 1$, the α-quantile, q_α, for a random variable X satisfies

$$P(X \leq q_\alpha) = \alpha.$$

Similarly, for $Z \sim N(0, 1)$, the α-quantile satisfies

$$P(Z \leq q_\alpha) = \Phi(q_\alpha) = \alpha.$$

This is illustrated in Figure 22.

For example, the upper quartile, q_U, satisfies

$$P(Z \leq q_U) = \Phi(q_U) = 0.75.$$

From Table 5,

$$P(Z \leq 0.67) = 0.7486, \quad P(Z \leq 0.68) = 0.7517.$$

So the value of q_U is between 0.67 and 0.68.

Alternatively, q_U can be found using Table 6. This table gives the value of q_α to four significant figures for various values of α from 0.5 to 0.999. This table is also included in the Handbook.

From Table 6, the upper quartile of Z is the value of q_α for $\alpha = 0.75$, that is, $q_{0.75} = 0.6745$. This is illustrated in Figure 23.

Similarly, the 0.9-quantile of Z is $q_{0.9} = 1.282$, and the 0.975-quantile of Z is $q_{0.975} = 1.960$.

Activity 16 *Practice at finding quantiles*

Use Table 6 to find the following quantiles of Z: $q_{0.6}$, $q_{0.85}$, $q_{0.99}$ and $q_{0.995}$.

Table 6 Standard normal quantiles

α	q_α	α	q_α	α	q_α	α	q_α
0.50	0.0	0.67	0.4399	0.84	0.9945	0.955	1.695
0.51	0.02507	0.68	0.4677	0.85	1.036	0.960	1.751
0.52	0.05015	0.69	0.4959	0.86	1.080	0.965	1.812
0.53	0.07527	0.70	0.5244	0.87	1.126	0.970	1.881
0.54	0.1004	0.71	0.5534	0.88	1.175	0.975	1.960
0.55	0.1257	0.72	0.5828	0.89	1.227	0.980	2.054
0.56	0.1510	0.73	0.6128	0.90	1.282	0.985	2.170
0.57	0.1764	0.74	0.6433	0.905	1.311	0.990	2.326
0.58	0.2019	0.75	0.6745	0.910	1.341	0.991	2.366
0.59	0.2275	0.76	0.7063	0.915	1.372	0.992	2.409
0.60	0.2533	0.77	0.7388	0.920	1.405	0.993	2.457
0.61	0.2793	0.78	0.7722	0.925	1.440	0.994	2.512
0.62	0.3055	0.79	0.8064	0.930	1.476	0.995	2.576
0.63	0.3319	0.80	0.8416	0.935	1.514	0.996	2.652
0.64	0.3585	0.81	0.8779	0.940	1.555	0.997	2.748
0.65	0.3853	0.82	0.9154	0.945	1.598	0.998	2.878
0.66	0.4125	0.83	0.9542	0.950	1.645	0.999	3.090

Table 6 shows q_α only for values $\alpha \geq 0.5$. The symmetry of the standard normal distribution allows Table 6 to also be used to find q_α for values $\alpha < 0.5$.

Consider q_α for $\alpha < 0.5$, as illustrated in Figure 24(a). Because $\alpha < 0.5$, q_α is negative. By the symmetry of $N(0, 1)$, the area under the curve to the left of $-q_\alpha$ ($-q_\alpha$ is a positive value) is $1 - \alpha$, as illustrated in Figure 24(b). Therefore $-q_\alpha$ must be the same as $q_{1-\alpha}$, or in other words, $q_\alpha = -q_{1-\alpha}$.

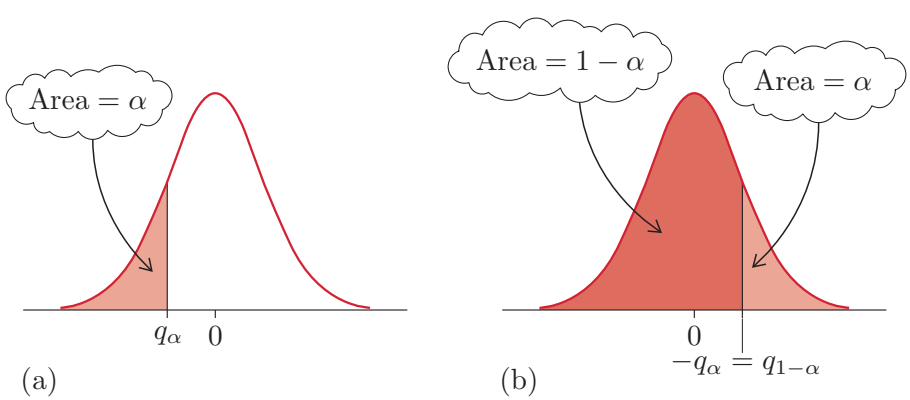


Figure 24 Using symmetry to find quantiles

Using symmetry to find quantiles

To find the quantile q_α for values $\alpha < 0.5$, use the result that

$$q_\alpha = -q_{1-\alpha}.$$

Example 16 *Using symmetry to find quantiles*

The lower quartile of Z is

$$q_{0.25} = -q_{1-0.25} = -q_{0.75} = -0.6745;$$

and the 30% percentage point, or 0.3-quantile, of Z is

$$q_{0.3} = -q_{1-0.3} = -q_{0.7} = -0.5244.$$

Activity 17 *Practice at finding quantiles using symmetry*

Use Table 6 and the symmetry of the standard normal curve to find the following quantiles of Z: $q_{0.4}$, $q_{0.2}$, $q_{0.05}$ and $q_{0.01}$.

Screencast 6.2 reviews finding quantiles of $N(0, 1)$ using Table 6.

Screencast 6.2 Using the standard normal table to find quantiles

4.4 Standardising normal random variables

This subsection explains how standard normal tables can be used to calculate probabilities and quantiles for *any* normal distribution – not just the standard normal distribution.

Recall from Distributional Result (3) that a linear function of a normal random variable is also a normal random variable. This result will be used to transform all normal distributions to the standard normal distribution.

A more extreme transformation

Consider random variable $X \sim N(\mu, \sigma^2)$. By Distributional Result (3) with $a = 1$ and $b = -\mu$, random variable $Y = X - \mu$ is normally distributed with

$$E(Y) = E(X) - \mu = \mu - \mu = 0 \quad \text{and} \quad V(Y) = 1^2 \times V(X) = \sigma^2,$$

that is,

$$Y = X - \mu \sim N(0, \sigma^2).$$

So $Y = X - \mu$ has the same mean as the standard normal distribution. This transformation is illustrated in Figure 25.

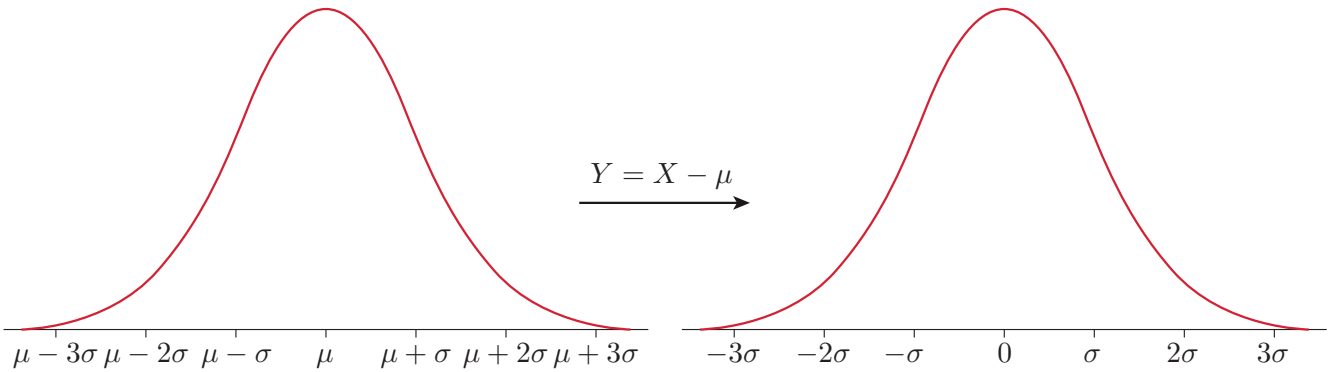


Figure 25 Transforming $X \sim N(\mu, \sigma^2)$ to $Y = X - \mu \sim N(0, \sigma^2)$

To complete the transformation to the standard normal random variable Z, random variable Y needs to be multiplied by $1/\sigma$ so that

$$Z = \frac{Y}{\sigma} = \frac{X - \mu}{\sigma}.$$

Then, by Distributional Result (3) with $a = 1/\sigma$ and $b = 0$, Z is normally distributed with

$$E(Z) = \frac{E(Y)}{\sigma} = \frac{0}{\sigma} = 0 \quad \text{and} \quad V(Z) = \frac{V(Y)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1,$$

that is, $Z \sim N(0, 1)$. The transformation from Y to Z is illustrated in Figure 26.

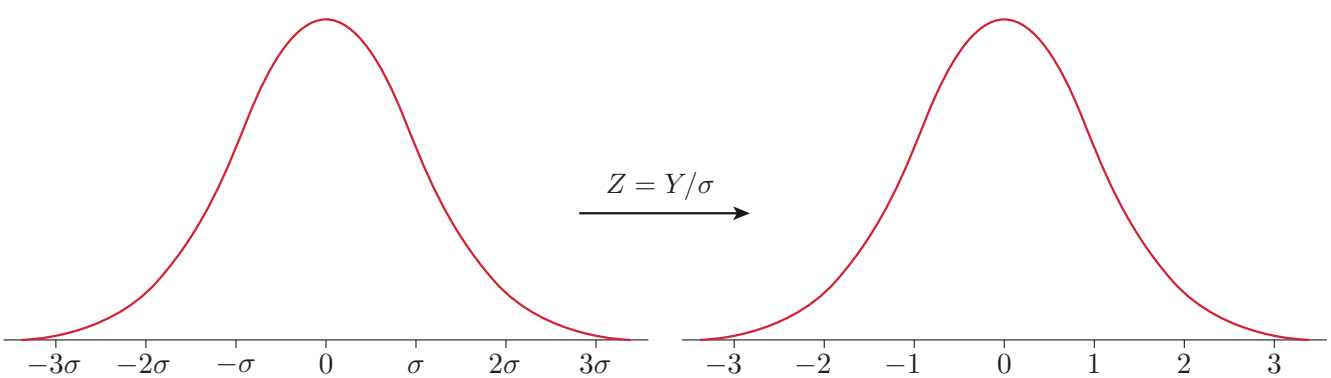


Figure 26 Transforming $Y \sim N(0, \sigma^2)$ to $Z = Y/\sigma \sim N(0, 1)$

Activity 18 *The inverse relationship*

We have just seen that if $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$. Conversely, if $Z \sim N(0, 1)$, how do you think X is related to Z so that $X \sim N(\mu, \sigma^2)$? Check your answer by applying Distributional Result (3) to your formula for X.

The procedure of subtracting the mean from a value of a random variable X and dividing by the standard deviation is known as **standardising** or **standardisation**. When a value of a normal random variable is standardised, the value obtained is an observation on the standard normal random variable Z. The relationship between X and Z is summarised in the box below.

Standardisation for normal distributions

If $X \sim N(\mu, \sigma^2)$, then the random variable

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1). \tag{11}$$

Conversely, if $Z \sim N(0, 1)$, then the random variable

$$X = \sigma Z + \mu \sim N(\mu, \sigma^2). \tag{12}$$

Z is said to be the standardised version of X.

This relationship between a general normal distribution and the standard normal distribution means that any problem about a normal random variable may be expressed as a problem about the standard normal random variable. Hence printed tables for the standard normal distribution may be used to calculate probabilities and quantiles for any normal distribution. Distributional Result (11) is required so often in statistics that you should try to memorise it.

In practice, to find a probability such as $P(X \leq x)$, where $X \sim N(\mu, \sigma^2)$, we rewrite it by standardising and using Distributional Result (11) as follows:

$$\begin{aligned} P(X \leq x) &= P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{x - \mu}{\sigma}\right). \end{aligned}$$

Then Table 5 can be used to find the required value.

This approach may be summarised as follows.

Calculating probabilities for normal distributions

If the random variable X has a normal distribution with mean μ and standard deviation σ, then

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right), \tag{13}$$

where Φ is the c.d.f. of the standard normal random variable Z.

The next two examples illustrate the application of the standardisation procedure summarised by the result in Equation (13).

Example 17 *Chest measurements revisited*

In Example 4 a normal distribution with parameters $\mu = 40$ and $\sigma = 2$ was proposed as a reasonable model for the chest measurements (in inches) of nineteenth-century Scottish soldiers represented in Figure 2. So, according to the model, the proportion of Scottish soldiers with chest measurements of at least 43 inches is given by

$$P(X \geq 43) = P\left(\frac{X - 40}{2} \geq \frac{43 - 40}{2}\right).$$

That is,

$$P(X \geq 43) = P\left(Z \geq \frac{43 - 40}{2}\right) = P(Z \geq 1.5).$$

This probability is illustrated in Figure 27. So

$$P(X \geq 43) = 1 - \Phi(1.5) = 1 - 0.9332 = 0.0668.$$

According to the model, fewer than 7% of Scottish soldiers had chest measurements of at least 43 inches.

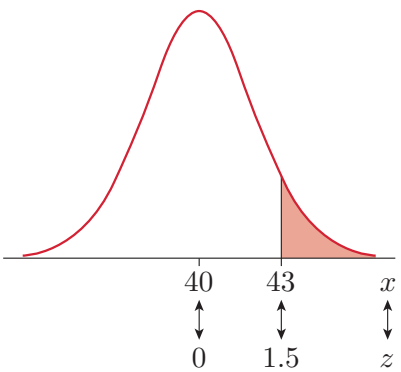


Figure 27 $P(X \geq 43) = P(Z \geq 1.5)$

Example 18 *IQ scores*

An IQ test is designed so that in the general population the variability in the scores attained should be normally distributed with mean 100 and standard deviation 15. (The variance is $15^2 = 225$.) This score is a random variable W, where the statistical model is $W \sim N(100, 225)$. (Most scales for measuring 'intelligence' are designed so that the mean is 100 and the standard deviation is 15.)

The proportion of scores between 80 and 120 is given by

$$P(80 \leq W \leq 120) = P\left(\frac{80 - 100}{15} \leq Z \leq \frac{120 - 100}{15}\right)$$
$$\simeq P(-1.33 \leq Z \leq 1.33).$$

This probability is illustrated in Figure 28. So the required probability is approximately equal to

$$\Phi(1.33) - \Phi(-1.33) = \Phi(1.33) - (1 - \Phi(1.33)) = 2\Phi(1.33) - 1$$
$$= 2 \times 0.9082 - 1 = 0.8164.$$

According to the model, just over 80% of the population have intelligence scores in the range 80 to 120.

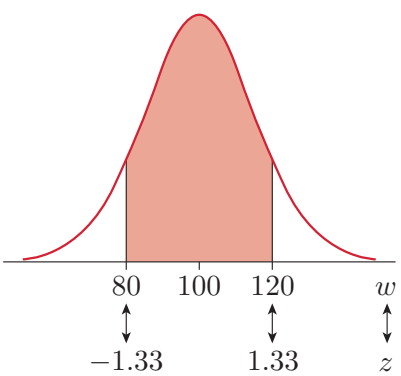


Figure 28 $P(80 \leq W \leq 120) = P(-1.33 \leq Z \leq 1.33)$

Note that a computer will give you the value 0.8176 for the probability in Example 18. The discrepancy between this value and the one obtained using tables is due to rounding error: in order to use Table 5, the number $20/15 = 4/3\,(= 1.333\ldots)$ was rounded to two decimal places. Consequently, although the values in the table are accurate to four decimal places, the value obtained for the probability in Example 18 is accurate to only two decimal places: it differs from the exact value by 1 in the third decimal place.

In most cases where rounding is involved in calculations, answers obtained using Table 5 will be accurate to three decimal places.

You will find that similar rounding is often necessary when using Table 5 to calculate probabilities. So you should not routinely claim four-decimal-place accuracy for your answers. We recommend that you use the full accuracy of the tables in your calculations, then round your answer to three decimal places. This is done in the units in all future calculations using Table 5.

Activity 19 *Heights of elderly women*

In Example 1, data were introduced concerning the heights (measured in cm) of 351 elderly women from the general population. Suppose that the following normal model is proposed for the distribution of H, the random variable representing the heights of the women: $H \sim N(160, 36)$. Use this model to calculate the following.

(a) The proportion of elderly women who are shorter than 150 cm in height.

(b) The proportion of elderly women who are between 155 cm and 165 cm tall.

Activity 20 *Bags of sugar*

On 5 February 2015, the *Washington Post* reported that India has the lowest average sugar consumption per person in the world, followed by Israel and then Indonesia

In Example 7, a situation was described in which a cook who requires 6 kg of sugar buys three bags, each of which is labelled as containing 2 kg. A normal model was assumed for X, the amount of sugar (in grams) in a bag: $X \sim N(2003, 10)$. In Example 9, you saw that the distribution of S, the total contents of three bags, has a normal distribution with mean 6009 and variance 30: $S \sim N(6009, 30)$.

What is the probability that the cook has less than 6 kg of sugar?

Activity 21 *Arriving late for work*

In Activity 10, a normal distribution, $N(15, 0.5)$, was used to model the distribution of the time (in minutes) that Jack takes to walk to the bus stop on his way to work. A normal distribution, $N(20, 9)$, was used to model the time (in minutes) from his arrival at the bus stop until he gets to work. Assuming that these times are independent, you found the distribution of the total time it takes Jack to get to work and wrote down an expression for the probability that he will arrive late if he leaves home 40 minutes before he is due at work. Find the value of this probability.

So far in this subsection, you have used Distributional Result (11) between a general normal distribution and the standard normal distribution to find probabilities associated with normal distributions. Now suppose that, for instance, we wish to know the IQ score which is exceeded by only 5% of the population. And what is the value x such that the chest measurements

of only 10% of nineteenth-century Scottish soldiers were smaller than x? Such problems are solved by finding quantiles of normal distributions.

You found quantiles for the standard normal distribution in Subsection 4.3. Quantiles for a normal distribution with mean μ and standard deviation σ may be calculated from quantiles for the standard normal distribution using Distributional Result (12): if $Z \sim N(0,1)$, then $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$. So if q_α is the α-quantile of $N(0,1)$, then

$$\alpha = P(Z \leq q_\alpha) = P\left(\frac{X-\mu}{\sigma} \leq q_\alpha\right) = P(X \leq \sigma q_\alpha + \mu).$$

It follows that the α-quantile of $N(\mu, \sigma^2)$, the distribution of X, is as given in the following box.

Finding quantiles of normal distributions

If q_α is the α-quantile of $N(0,1)$, then the α-quantile, x, of $N(\mu, \sigma^2)$ is given by

$$x = \sigma q_\alpha + \mu. \tag{14}$$

Example 19 *IQ scores*

The IQ score, x, which is exceeded by only 5% of the population is the 0.95-quantile of $N(100, 225)$. From Table 6, the 0.95-quantile of $N(0,1)$ is $q_{0.95} = 1.645$. So, using Equation (14), the 0.95-quantile of $N(100, 225)$ is given by

$$x = 15\, q_{0.95} + 100 = 15 \times 1.645 + 100 \simeq 124.7.$$

So, according to the model, the IQ scores of only 5% of the population are greater than 124.7.

Example 20 *Chest measurements*

The value x such that the chest measurements of only 10% of nineteenth-century Scottish soldiers were smaller than x is the 0.1-quantile of $N(40, 4)$. The 0.1-quantile of $N(0,1)$ is $q_{0.1} = -q_{0.9} = -1.282$. So the 0.1-quantile of $N(40, 4)$ is given by

$$x = 2\, q_{0.1} + 40 = 2 \times (-1.282) + 40 \simeq 37.4.$$

So, according to the model, the chest measurements of only 10% of nineteenth-century Scottish soldiers were less than 37.4 inches.

Activity 22 *IQ scores*

Assuming that the normal distribution $N(100, 225)$ is an adequate model for IQ scores, find the quantiles $q_{0.2}$, $q_{0.4}$, $q_{0.6}$, $q_{0.8}$ for IQ scores. Illustrate these quantiles on a sketch of the p.d.f. of $N(100, 225)$.

Octogenarian Dilys Price OBE, founder of the Touch Trust charity for profoundly disabled people and winner of the Lifetime Achiever Award at the UK's National Diversity Awards in 2014, is also a prolific skydiver and has been registered as the world's oldest female skydiver by Guinness World Records

Activity 23 *Heights of elderly women*

A normal model with mean 160 and standard deviation 6 has been proposed for the distribution of the heights (in centimetres) of elderly women in the general population.

(a) According to the model, what is the median height of elderly women in the population?

(b) Find the height such that only 1% of elderly women are taller than this height.

(c) Find the height such that only 15% of elderly women are shorter than this height.

The final activity in this section involves finding both probabilities and quantiles for a normal distribution.

Nicotine is named after Jean Nicot de Villemain (1530–1600) who introduced tobacco to the French royal court, promoting it for medicinal use!

Activity 24 *Nicotine levels of smokers*

Suppose that blood plasma nicotine levels in smokers may be modelled by a random variable T which is normally distributed with mean 315 and standard deviation 131 ng/ml (that is, nanograms per millilitre).

(a) According to the model, what proportion of smokers have nicotine levels above 450?

(b) In practice, a negative nicotine level is impossible. According to the model, what proportion of smokers have nicotine levels below zero? Comment on the adequacy of the model.

(c) According to the model, what is the interquartile range of nicotine levels?

(d) What nicotine level is such that the nicotine levels of only 4% of smokers are higher?

Exercises on Section 4

Exercise 6 *Practice at using printed tables*

For each of the following probabilities, draw a rough sketch of the p.d.f. of the standard normal distribution and mark on your sketch the area which represents the probability. Then use Table 5 (or the table in the Handbook) to find the value of the probability.

(a) $P(Z > -0.42)$

(b) $P(0.17 < Z < 1.17)$

(c) $P(|Z| \leq 2.25)$

(d) $P(|Z| \geq 1.75)$

Exercise 7 *Practice at finding quantiles of Z*

Use Table 6 (or the table of quantiles in the Handbook) to find the following quantiles of Z: $q_{0.998}$, $q_{0.55}$, $q_{0.35}$, $q_{0.1}$ and $q_{0.001}$.

Exercise 8 *Heights of men*

Suppose that an adequate model for the heights (in centimetres) of men in a population is a normal distribution, $N(173, 40)$.

(a) According to the model, what proportion of men in the population are shorter than 160 cm?

(b) According to the model, what proportion of men in the population are between 170 cm and 180 cm tall?

(c) Find the height x such that only 2% of men in the population are taller than this height.

(d) Find the interquartile range of the distribution of heights of men in the population.

Exercise 9 *Bags of flour*

Suppose that the contents (in grams) of bags of flour, each labelled as containing 1.5 kg, may be adequately modelled by a normal distribution, $N(1501, 2.5)$. Find the probability that the total contents of ten randomly selected bags of flour will be less than 15 kg.

5 Normal probability plots

For each of the datasets represented in Figures 1, 2 and 3, it was observed that a histogram of the data is essentially unimodal and roughly symmetric about its peak. Since the p.d.f. of a normal distribution has these properties, a normal model was proposed for the variation in each dataset. For these histograms, the datasets are large and the fits look fairly good; see Figure 6. However, for other datasets, particularly small ones, it can be less clear that a normal model really is a good fit for the data. In this section, an alternative graphical method for investigating whether a normal model is a good fit for data is described. These plots are called **normal probability plots**.

Suppose that you have n observations, $x_1, x_2, \ldots, x_n$, and that you wish to know whether they may plausibly have arisen from a normal distribution. A normal probability plot for the data is obtained as follows.

Unit 1 also looks at ordering data using this notation.

- First, rearrange the data into order of increasing size. If the ith ordered observation is denoted by $x_{(i)}$, then

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

- Next calculate the n quantiles, $y_1, y_2, \ldots, y_n$, for a standard normal distribution:

$$y_i = q_{i/(n+1)}, \quad i = 1, 2, \ldots, n.$$

  The values $y_1, y_2, \ldots, y_n$ are known as **normal scores**.

- Finally, plot the n points $(x_{(i)}, y_i)$, $i = 1, 2, \ldots, n$, on a graph. This graph is the normal probability plot for the data.

In the normal scores, the subscript of q is $i/(n+1)$ not i/n; if it were i/n, we'd have $y_n = q_{n/n} = q_1$, which would be ∞.

If a normal distribution is a plausible model for the data, then the points will lie approximately on a straight line; Screencast 6.3 will explain why this is so.

Screencast 6.3 Why normal probability plots work

Normal probability plots

If the normal distribution is a good fit to the data, then the points in a normal probability plot will lie roughly along a straight line.

Note that you do not need to specify values for the parameters of the normal distribution to obtain a probability plot. (In fact, if the points lie roughly along a straight line, then it is possible to use the slope and intercept of the line to calculate estimates for the parameters. However, we will not discuss how to do this in this module.) If you assess normality using a histogram and a superimposed normal p.d.f., as in Figure 6, not only do you need to specify values for the parameters of the normal distribution, you also need to choose the starting position of the bins of the histogram and their width, as discussed in Unit 1. You need make none of these choices to use a normal probability plot.

Example 21 *Silver content of Byzantine coins*

The silver content (% Ag) of a number of Byzantine coins discovered in Cyprus was determined. Nine of the coins came from the first coinage of the reign of Manuel I Comnenus (1143–1180), there were seven from the second coinage minted several years later, four from the third coinage (later still), and another seven were from the fourth coinage. The question of interest is whether there were differences in the silver content of coins minted early and late in Manuel's reign. One method of investigating this question depends on the assumption that the variation in the silver content of coins from each coinage may be modelled adequately by a normal distribution. For a small dataset, it is not possible to judge whether a normal model is a reasonable one simply by producing a histogram of the data. This is the sort of situation where a normal probability plot is more useful.

The construction of a normal probability plot will be illustrated for the data on the silver content of coins from the first coinage. The data are in Table 7.

Manuel I Comnenus

Table 7 Silver content of coins from the first coinage (% Ag)

5.9	6.8	6.4	7.0	6.6	7.7	7.2	6.9	6.2

(Source: Hendy, M.F. and Charles, J.A. (1970) 'The production techniques, silver content and circulation history of the twelfth-century Byzantine trachy', *Archaeometry*, vol. 12, no. 1, pp. 13–21)

First, the data are listed in order of increasing size. Then, since the sample size n is 9, the quantiles required are $y_1, y_2, \ldots, y_9$, where y_i is the quantile $q_{i/10}$ of the standard normal distribution. That is, we need to find $q_{0.1}, q_{0.2}, \ldots, q_{0.9}$. The results are shown in Table 8.

Table 8 The points $(x_{(i)}, y_i)$

i	$x_{(i)}$	$i/10$	y_i
1	5.9	0.1	−1.282
2	6.2	0.2	−0.842
3	6.4	0.3	−0.524
4	6.6	0.4	−0.253
5	6.8	0.5	0.000
6	6.9	0.6	0.253
7	7.0	0.7	0.524
8	7.2	0.8	0.842
9	7.7	0.9	1.282

A normal probability plot is shown in Figure 29.

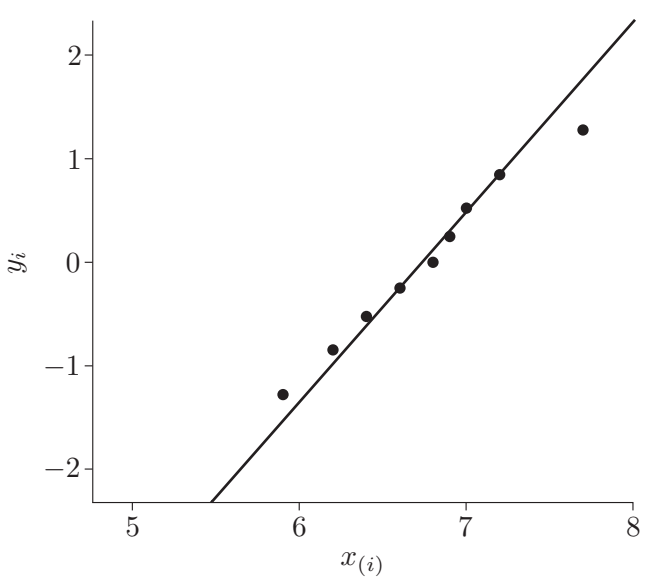


Figure 29 A normal probability plot for silver content for the first coinage

As you can see, the points do lie approximately along a straight line. So it looks as though a normal distribution is a plausible model for the silver content of coins from the first coinage.

Since producing a probability plot involves a lot of calculation, it is a task ideally suited to the computer. So all the activities associated with this section are in Computer Book B.

Refer to Chapter 2 of Computer Book B for the rest of the work in this section.

6 The sampling distribution of the mean

If, on two different occasions, you were to draw a random sample of the same size from the same population, then you would almost certainly get different individual observations and different sample means and variances. For instance, suppose that the heights of a random sample of 50 ten-year-old girls are measured. The sample mean provides an estimate of the average height of all ten-year-old girls (the population mean). However, if a different sample of 50 ten-year-old girls were to be taken, this sample would almost certainly lead to different results and the sample mean would almost certainly be different.

In general, the sample mean of samples of size n will vary from sample to sample for a given population. Since the sample mean varies from sample to sample, *the sample mean is itself a random variable* and, as such, it has a distribution. The distribution of sample means is called the **sampling distribution of the mean**.

In this section, the sampling distribution of the mean will be explored for sample means obtained from different distributions. You will discover – or be reminded – that normal distributions have an important role to play.

Screencast 6.4 demonstrates what the sampling distribution of the mean looks like when the population distribution is normal.

Screencast 6.4 The sampling distribution of the mean of a normal distribution

In Screencast 6.4, you saw how when samples are taken from a population with a normal distribution, it looks like the sampling distribution of the mean is symmetric about a peak at μ and might also be normal. You also saw that as the size of the sample increases, so the spread of the sample means around the peak at μ decreases.

It is perhaps not a huge surprise that when samples are drawn from a population with a normal distribution, the sampling distribution of the mean looks normal. But what does the sampling distribution of the mean look like when samples are taken from a population which is not normally distributed? This is explored in Screencast 6.5.

Screencast 6.5 The sampling distribution of the mean of non-normal distributions

In Screencast 6.5, you saw how when samples are taken from populations with distributions other than the normal distribution, and when the sample sizes are large, it looks like the sampling distribution of the mean is symmetric about a peak at μ and might also be normal. In fact, for sufficiently large samples, the sampling distribution of the mean *is* approximately normal, no matter what the underlying population distribution is. The population can be very non-normal; it can even be discrete! This is an important result in statistics and is part of what is known as the *Central Limit Theorem*. (A formal statement of the Central Limit Theorem will be given in Subsection 6.2.)

The sampling distribution of the mean

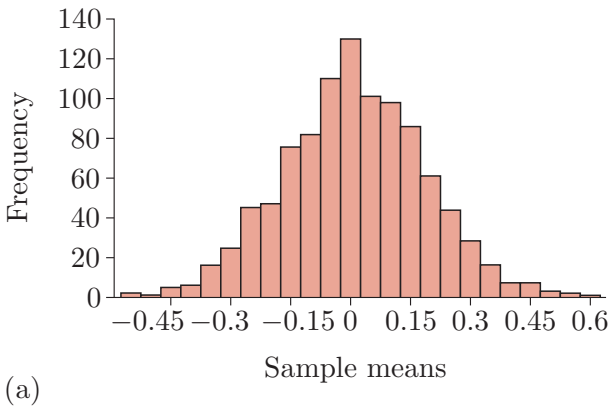
Whatever the distribution of the population, for sufficiently large samples, the sample mean is approximately normally distributed.

Actually, there are exceptions, but they are few and not usually of practical importance

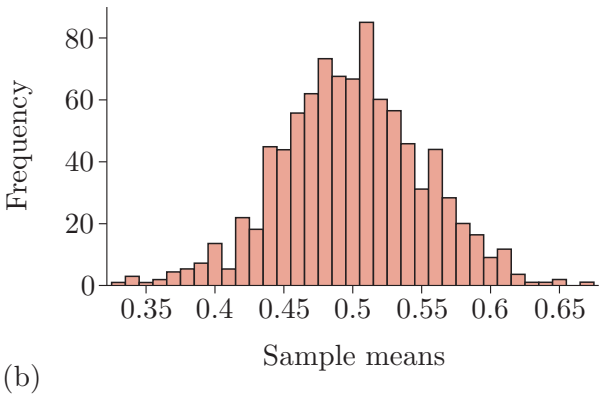
This result is illustrated in Figure 30 (overleaf).

Notice that the *sampling* distribution (of the sample mean) is itself a *population* distribution (and not a *sample* distribution). This is because, as we said earlier, the sample mean is itself a random variable and, as such, has a population distribution. But, as you have already seen, the sampling distribution of the sample mean is a population distribution that, in general, differs from the population distribution underlying the individual sample values.

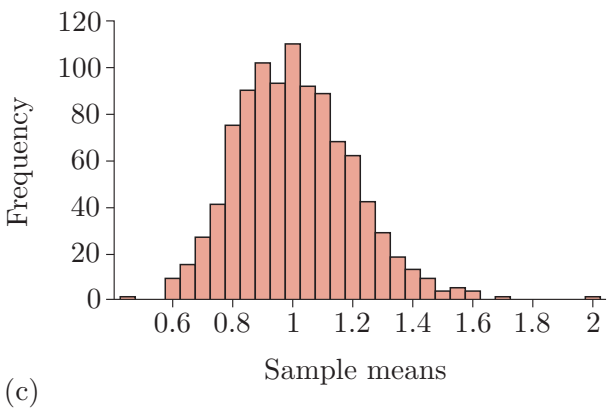
In order to use the Central Limit Theorem, the mean and variance of the approximate normal distribution for the sampling distribution of the mean are required. This is the subject of the next subsection.



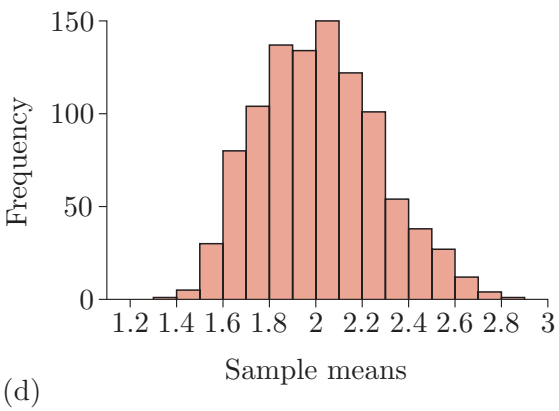
(a)



(b)



(c)



(d)

Figure 30 The sampling distributions of means of samples of size $n = 30$ from (a) $N(0, 1)$, (b) $U(0, 1)$, (c) $M(1)$, (d) Poisson(2) distributions; all appear approximately normally distributed

6.1 The sample total and the sample mean

Suppose that we have a random sample of size n from a population and that the values in the sample are $x_1, x_2, \ldots, x_n$. The *sample total* is simply the sum of all the values in the dataset:

$$t_n = x_1 + x_2 + \cdots + x_n.$$

The *sample mean* is the sample total divided by the sample size:

$$\overline{x}_n = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\, t_n.$$

Notice that the subscript n has been included in the notation used here for the sample total and the sample mean. This has been done so that the size of the sample is explicit in what follows. For samples of the same size n drawn from the same population, we would expect to observe variability in the individual data values and also in the sample total and the sample mean.

Therefore, in any single experiment, the sample total t_n is just one observation on a random variable T_n, and the sample mean $\overline{x}_n$ is just one observation on a random variable $\overline{X}_n$. If X is a random variable representing the attribute of interest in the population, then

$$T_n = X_1 + X_2 + \cdots + X_n \tag{15}$$

and

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}T_n, \tag{16}$$

where $X_1, X_2, \ldots, X_n$ are independent observations on X.

The convention that upper-case letters denote random variables and lower-case letters denote observations thereon is being used here.

You can now work out the expected values of the sample total T_n and the sample mean $\overline{X}_n$ in Activity 25, and then their variances in Activity 26. Note that for a random sample, the components in the sample total T_n, as well as being identically distributed, are also independent.

Activity 25 *The expected value of the sample total and the sample mean*

Suppose that a random sample of size n is drawn from a population with mean μ, so that $E(X) = \mu$, where X is the attribute of interest in the population. Use Equations (4) and (1), respectively, to answer the following.

(a) Find an expression for the expected value of the sample total T_n, where T_n is defined by Equation (15).

(b) Find an expression for the expected value of the sample mean $\overline{X}_n$, where $\overline{X}_n$ is defined by Equation (16).

Activity 26 *The variance of the sample total and the sample mean*

Suppose that a random sample of size n is drawn from a population with variance σ^2, so that $V(X) = \sigma^2$, where X is the attribute of interest in the population. Use Equations (5) and (2), respectively, to answer the following.

(a) Find an expression for the variance of the sample total T_n, where T_n is defined by Equation (15).

(b) Find an expression for the variance of the sample mean $\overline{X}_n$, where $\overline{X}_n$ is defined by Equation (16).

The results obtained in Activities 25 and 26 are important. They may be summarised as follows.

Properties of the sample total and the sample mean

If X is a random variable with mean μ and variance σ^2, and a random sample of size n is taken from the distribution of X, then the mean and variance of the sample total are given by

$$E(T_n) = n\mu, \tag{17}$$

$$V(T_n) = n\sigma^2, \tag{18}$$

and the mean and variance of the sample mean are given by

$$E(\overline{X}_n) = \mu, \tag{19}$$

$$V(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{20}$$

Equation (19) is very reassuring: it states that the expected value of the sample mean is equal to the population mean μ. This confirms the findings of Screencasts 6.4 and 6.5 in which the sampling distribution of the mean appeared to be centred at the population mean, as well as the intuitive view that the sample mean should, at least in some sense, be a good indicator or predictor (or, using the most usual statistical terminology, a good estimator) of the unknown population mean.

Note that Equation (19) does not depend on the value of n, the sample size. In a sample of size 2, the sample mean has expected value μ; and in a sample of size 100, or any other size, the sample mean also has expected value μ.

Notice that the sample size n *does* feature in Equation (20): the larger the sample size, the smaller the variance of the sample mean. This confirms the observation in Screencasts 6.4 and 6.5 that the spread of the sampling distribution of the mean decreases as the sample size increases. A sample mean obtained from a large sample is therefore a more reliable estimator of the unknown population mean μ than a sample mean obtained from a smaller sample, in the sense that it is less likely to be very far away from its expected value, which in both cases is the unknown population mean μ. This result confirms the intuition that larger samples are 'better' than smaller samples.

Loosely speaking, in a notional sample of size ∞, the sample mean would be exactly the population mean. Then, there's no variability in the sample mean: its expected value is μ and its variance is zero.

The use of Equations (17) and (18) for the sample total is illustrated in Example 22.

Example 22 *Time for coffee*

Suppose that the duration of a patient's visit to a dentist's surgery is a random variable X with mean 10 minutes and standard deviation 5 minutes. The dentist attends to eight patients each morning before taking a coffee break.

If $X_1, X_2, \ldots, X_8$ represent the times that the dentist spends with each of her first eight patients, then the total time T_8 that she works before taking a coffee break is given by

$$T_8 = X_1 + X_2 + \cdots + X_8.$$

So, applying Equation (17), the expected time (in minutes) until she takes a coffee break is

$$E(T_8) = 8E(X) = 8 \times 10 = 80;$$

and, using Equation (18), the variance of this time is

$$V(T_8) = 8V(X).$$

The standard deviation of X is 5 minutes. The variance is the square of the standard deviation, so, in units of minutes squared,

$$V(T_8) = 8 \times 25 = 200.$$

Hence the standard deviation of the time (in minutes) that the dentist works before taking a coffee break is $\sqrt{200}$, or approximately 14.1 minutes.

In the activity below, you will also need to use Equations (19) and (20) for the mean and variance of the sample mean.

Activity 27 *Mean weight of bags of sugar*

Suppose that the mean contents of bags of sugar labelled as containing 1 kg is 1003 g, and that the standard deviation of their contents is 2 g. A random sample of five bags is taken.

(a) Find the expected value and the standard deviation of the total contents of a sample of five bags.

(b) Find the expected value and the standard deviation of the mean contents of a sample of five bags.

Plenty of sugar was required for this cake baked for the University of Chicago Department of Statistics to celebrate the 275th birthday of the normal distribution in 2008

6.2 The Central Limit Theorem

We are now in a position to formally state the Central Limit Theorem. The proof of the theorem requires advanced mathematical analysis, so the theorem is stated here without proof.

The Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are n independent random observations from a population with mean μ and finite variance σ^2, then for large n the distribution of their mean $\overline{X}_n$ is approximately normal with mean μ and variance σ^2/n:

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right). \qquad (21)$$

Recall that the symbol '$\approx$' is read 'has approximately the same distribution as'.

The Central Limit Theorem states that for large sample sizes, the sampling distribution of the mean is approximately normal with mean μ

and variance σ^2/n. The standard deviation of the sampling distribution of the mean, which is equal to $\sigma/\sqrt{n}$, is frequently called the **standard error** of the mean.

This is standard terminology: 'error' should not be read as 'mistake'.

Note that the Central Limit Theorem is an asymptotic result; that is, the approximation improves as the sample size increases. The theorem allows us to make a number of statements about the usefulness of the sample mean as an estimator of the population mean.

- If a random sample is drawn from a population with unknown mean μ, then since $E(\overline{X}_n) = \mu$, the sample mean is a 'good' indicator of the unknown population mean μ.
- Since the variance of the sample mean decreases as n increases ($V(\overline{X}_n) = \sigma^2/n$), the larger the sample that is taken, the more likely it is that the sample mean will be close to the population mean μ, and so the more reliance we can place on the sample mean as an estimator of μ.
- Provided that the samples are sufficiently large, the symmetry of the normal distribution implies that in repeated experiments, the sample mean is as likely to underestimate as to overestimate the population mean μ; and this is so whatever the shape of the underlying population.

An important practical question is: 'How large must n be for the Central Limit Theorem to hold?' There is no precise answer to this question, but a general rule of thumb is to take n to be at least 25.

For many populations, n can be much less than 25.

The use of the Central Limit Theorem is illustrated in Example 23.

Example 23 *Mean height of meerkats*

Suppose that the mean standing height of adult meerkats is 30 cm, and the standard deviation of their heights is 1 cm. If a sample of 50 adult meerkats is taken, what is the probability that their mean standing height, $\overline{X}_{50}$, will be within 0.25 cm of the mean standing height of the population of adult meerkats?

A sample of size 50 is fairly large so, by the Central Limit Theorem, the distribution of the sample mean $\overline{X}_{50}$ is approximately normal. Moreover, its mean is equal to the population mean, that is, 30, and its variance is $\sigma^2/n = 1^2/50 = 0.02$. So $\overline{X}_{50} \approx N(30, 0.02)$. Hence the probability that the mean standing height of the sample of adult meerkats will be within 0.25 cm of the population mean is approximately given by

$$P(29.75 < \overline{X}_{50} < 30.25) \simeq P\left(\frac{29.75 - 30}{\sqrt{0.02}} < Z < \frac{30.25 - 30}{\sqrt{0.02}}\right)$$

$$\begin{aligned} &\simeq P(-1.77 < Z < 1.77) \\ &= \Phi(1.77) - \Phi(-1.77) \\ &= \Phi(1.77) - (1 - \Phi(1.77)) \\ &= 2\Phi(1.77) - 1 = 2 \times 0.9616 - 1 = 0.9232. \end{aligned}$$

So, for more than 92% of samples of size 50, the mean standing height will differ from the population mean by less than 0.25 cm.

Activity 28 *Mean height of meerkats*

For the population of meerkats in Example 23, find the approximate probability that the mean standing height of a sample of 40 adult meerkats will be less than 29.75 cm.

Activity 29 *Sample mean from the geometric distribution*

Suppose that a random sample of size $n = 100$ is taken from the geometric distribution with mean p. As you saw in Unit 4, the mean and variance of individual observations from this distribution are $1/p$ and $(1-p)/p^2$, respectively. What is the approximate distribution of the sample mean of this random sample?

You will see further use being made of the Central Limit Theorem as you work through other units in M248.

6.3 A corollary to the Central Limit Theorem

In this subsection, a corollary to the Central Limit Theorem is considered briefly. This concerns the sample total T_n.

Distributional Result (3) states that if the random variable X is normally distributed and a and b are constants, then the random variable $aX + b$ is also normally distributed. So, in particular, a constant multiple of a normally distributed random variable is also normally distributed. Now $T_n = n\overline{X}_n$, which is a constant multiple of $\overline{X}_n$, and by the Central Limit Theorem the sample mean $\overline{X}_n$ is approximately normally distributed. It follows that the distribution of T_n is also approximately normal. This result for the sample total may be stated formally as follows.

A corollary to the Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are n independent random observations from a population with mean μ and finite variance σ^2, then for large n the distribution of their sum T_n is approximately normal with mean $n\mu$ and variance $n\sigma^2$:

$$T_n = X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2). \qquad (22)$$

Example 24 illustrates the use of this corollary.

Example 24 *Counting traffic*

Traffic counts are still performed manually as well as by a variety of automatic methods

Vehicles pass an observer in such a way that the waiting time between successive vehicles may be adequately modelled by an exponential distribution with mean 15 seconds. Particular details of each vehicle that passes are recorded on a sheet. There is room to record the details of twenty vehicles on each sheet.

What, approximately, is the probability that it takes less than 6 minutes to fill one of the sheets?

If W is the waiting time in seconds between successive vehicles, then W has mean 15. Since the mean and standard deviation of an exponential distribution are equal, the variance of W is $15^2 = 225$. The time T_{20} that it takes to fill a sheet is the sum of the first twenty such waiting times:

$$T_{20} = W_1 + W_2 + \cdots + W_{20}.$$

So

$$E(T_{20}) = 15 + 15 + \cdots + 15 = 20 \times 15 = 300;$$

and, assuming that the times are independent,

$$V(T_{20}) = 225 + 225 + \cdots + 225 = 20 \times 225 = 4500.$$

Also, by the corollary to the Central Limit Theorem, Distributional Result (22), the distribution of T_{20} is approximately normal, so

$$T_{20} \approx N(300, 4500).$$

Hence the approximate probability that the total time taken to fill a sheet is less than 6 minutes (or 360 seconds) is given by

$$P(T_{20} < 360) \simeq P\left(Z < \frac{360 - 300}{\sqrt{4500}}\right) \simeq P(Z < 0.89) = 0.8133 \simeq 0.813.$$

In fact, if the exact distribution of the sample total is used instead of the normal approximation, this yields the value 0.8197. (You are not expected to be able to do the exact calculation.)

On 3 June 2009, *The Telegraph* reported that 24.6% of British adults will avoid opening bank statements at all costs, and 59% don't check bank balances until trouble hits

Activity 30 *Estimation errors*

Rather than keep an accurate record of individual transactions, the holder of a bank account records only individual deposits into and withdrawals from his account to the nearest pound. Suppose that the error in individual records may be adequately modelled by a continuous uniform distribution $U\left(-\frac{1}{2}, \frac{1}{2}\right)$ and that he makes 400 transactions in a particular year.

(a) What, approximately, is the distribution of the error in his estimate of his bank balance at the end of the year? Hint: it will be useful to remember that for $X \sim U(a, b)$, $E(X) = (a + b)/2$ and $V(X) = (b - a)^2/12$.

(b) Find the probability that the error in his estimate is less than £10.

Notice that in each of Example 24 and Activity 30, the continuous random variable T_n has been approximated by a normal random variable with the same mean and variance as T_n. In general, using a normal approximation for a continuous random variable X involves using a normal distribution with the same mean and variance as X to calculate approximate values for probabilities involving X.

6.4 Normal population distribution

The Central Limit Theorem says that the distribution of the sample mean is approximately normal for sufficiently large samples, regardless of the population distribution. However, in Screencast 6.4, you saw that when the population distribution is normal, the sampling distribution of the mean looks approximately normal, even for very small sample sizes. In fact, when the population distribution is normal, the normality of the sampling distribution of the mean is *exact*, rather than approximate, regardless of the sample size.

To see this, recall two results.

- If the random variable X is normally distributed and a and b are constants, then the random variable $Y = aX + b$ is also normally distributed. See Distributional Result (3).
- If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables, then their sum is also normally distributed. See Distributional Result (6).

It follows from the second result that if a random sample $X_1, X_2, \ldots, X_n$ is drawn from a population which is normally distributed, then the sample total T_n is normally distributed. And since $\overline{X}_n = T_n/n$, by the first result, $\overline{X}_n$ is also normally distributed. So for the case of sampling from a normal distribution, the sample total and the sample mean are normally distributed. Thus we have the following result.

Samples from a normal distribution

For random samples of size n from a normal distribution with mean μ and variance σ^2, the sample total T_n and the sample mean $\overline{X}_n$ are normally distributed:

$$T_n \sim N(n\mu, n\sigma^2), \quad \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

It is worth stressing again that the difference between these results and the similar-looking ones in Subsections 6.2 and 6.3 is that these results are exact because the underlying population is normal, while the other results are approximate because the underlying population is not normal.

Coffee break at last!

Example 25 *A late coffee break*

Suppose that the dentist in Example 22 starts work at 9 a.m. and that the time (in minutes) that she spends with each patient may be adequately modelled by a normal distribution with mean 10 and standard deviation 5. Then T_8, the time that she works before taking a coffee break after seeing eight patients, is also normally distributed: $T_8 \sim N(80, 200)$. So the probability that it is after 10:30 a.m. when she begins her coffee break is given by

$$P(T_8 > 90) = P\left(Z > \frac{90 - 80}{\sqrt{200}}\right) \simeq P(Z > 0.71)$$
$$= 1 - \Phi(0.71) = 1 - 0.7611 = 0.2389 \simeq 0.239.$$

Activity 31 *Bags of sugar*

In Activity 27, you found the mean and standard deviation of the sampling distribution of the mean for samples of five bags of sugar from a population with mean 1003 g and standard deviation 2 g. Assuming that the contents of the bags of sugar are normally distributed, find the probability that the mean contents of a sample of five bags will be less than 1 kg.

Exercises on Section 6

Exercise 10 *Van carrying capacity*

A van has a weight carrying capacity, or 'payload', of 1600 kg. The company that owns the van wishes to transport 30 small but heavy items which it knows from long experience have a mean weight of 50 kg and a standard deviation of 10 kg. What is the approximate probability that the total weight of the 30 items will be less than the payload of the van?

Exercise 11 *Bottles of orange juice*

Suppose that the volume of orange juice in bottles, each of which is labelled as containing 1 litre, is normally distributed with mean 1004 ml and standard deviation 5 ml.

(a) Find the probability that the total contents of three randomly selected bottles will be less than 3 litres.

(b) Find the probability that the mean contents of a random sample of four bottles will be less than 1 litre.

Summary

A normal distribution is a continuous distribution and is useful for modelling the variation observed in a range of phenomena. In this unit, normal distributions have been discussed in detail.

An important property of a normal distribution is that the area under a graph of its p.d.f., within k standard deviations of the mean, depends only on the value of k and not on the values of the mean μ and the standard deviation σ. This means that any problem involving a normal distribution may be expressed as a problem about the normal distribution with mean 0 and standard deviation 1 – the standard normal distribution. This is done using a procedure known as standardisation. You have learned how to use printed tables for the standard normal distribution to find probabilities and quantiles for any normal distribution.

You have seen that when the random variable X is normally distributed, a linear function of X is also normally distributed. You have also seen that a sum of independent normal random variables has a normal distribution.

Normal probability plots have been introduced. Obtaining a normal probability plot for a set of data involves plotting the ordered data values against particular quantiles of the standard normal distribution (the normal scores). If a normal model is a good fit for the data, then the plotted points will lie roughly along a straight line.

You have used Minitab to find probabilities and quantiles for normal distributions and to obtain normal probability plots.

Finally, one of the fundamental results of statistical theory, the Central Limit Theorem, has been discussed. The theorem states that for large sample sizes, the sampling distribution of the mean for samples of size n from a population with mean μ and variance σ^2 is approximately normal with mean μ and variance σ^2/n. A corollary to this theorem states that the distribution of the sample total has mean $n\mu$ and variance $n\sigma^2$, and is also approximately normal for large samples. You will meet further applications of the Central Limit Theorem in later units of M248.

Learning outcomes

After you have worked through this unit, you should be able to:

- remember that the parameters μ and σ of a normal distribution are the mean and standard deviation of the distribution
- appreciate that if a normal distribution is used to model the variation in a population, then according to the model, the proportion of the population within k standard deviations of the mean is independent of the parameters of the normal distribution
- understand that both a linear function of a normal random variable and a sum of independent normal random variables follow a normal distribution
- appreciate that if $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$, and that if $Z \sim N(0, 1)$, then $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$
- recognise the normal distribution with mean 0 and variance 1 as the standard normal distribution, for which standard notation includes Z for the standard normal random variable, and ϕ and Φ for its p.d.f. and c.d.f., respectively
- use tables to find probabilities and quantiles for the standard normal distribution
- use standardisation and tables for the standard normal distribution to solve problems involving a normal distribution with mean μ and standard deviation σ
- use Minitab to find probabilities and quantiles for normal distributions
- use normal probability plots to assess whether or not a normal model is a good fit for a sample of data: if a normal distribution is a good fit, then the points on such a plot should lie roughly along a straight line
- use Minitab to obtain a normal probability plot for a dataset
- appreciate that the sample total and the sample mean in repeated experiments are random variables
- calculate the mean and variance of the sampling distribution of the mean for samples of size n from a population with mean μ and variance σ^2, namely μ and σ^2/n, respectively
- calculate the mean and variance of the sampling distribution of the sample total for samples of size n from a population with mean μ and variance σ^2, namely $n\mu$ and $n\sigma^2$, respectively
- understand that when large samples are drawn from a population, the distributions of the sample mean and the sample total are approximately normal, whatever the distribution of the population; these results are the Central Limit Theorem and its corollary; the approximation improves as the sample size increases; also, the distributions are exactly normal when the population is normal
- use the Central Limit Theorem and its corollary to solve problems involving the sample total or the sample mean when the sample size is large.

Solutions to activities

Solution to Activity 1

Ignoring any jaggedness which might simply be due to random variation, both of the histograms may be described as unimodal and roughly symmetric about a central peak. In each case, frequencies are small for values on the left of the diagram, increase fairly steadily, reaching a maximum in the centre, then decrease at the same rate as they increased, and are small for values on the right. (This description applies immediately to Figure 2. It also applies to Figure 1, provided that the two largest apparent peaks towards the centre of the histogram are ascribed to random variation and to the use of a larger number of intervals than in Figure 2.)

Solution to Activity 2

The two histograms have the same basic shape: they are unimodal and roughly symmetric about a central peak. The positions of the two peaks do not differ greatly. However, although the average enzyme measurements do not differ greatly, the boxplots in Figure 4 suggest that the measurements on the liver enzyme are slightly more variable for the group of patients suffering from aggressive chronic hepatitis than for those suffering from acute viral hepatitis.

Solution to Activity 3

All normal curves are bell-shaped and symmetric about a single peak centred at μ. In addition, almost all the distribution lies between the values $\mu \pm 3\sigma$.

(a) When $\mu = 10$ and $\sigma = 5$, the normal curve is symmetric about 10 and lies between the values $10 \pm (3 \times 5)$, that is, -5 and 25. In particular, the labels in Figure 7 can be replaced by $\mu - 3\sigma = 10 - 3 \times 5 = -5$, $\mu - 2\sigma = 10 - 2 \times 5 = 0$, and so forth. The normal curve is illustrated in Figure 31.

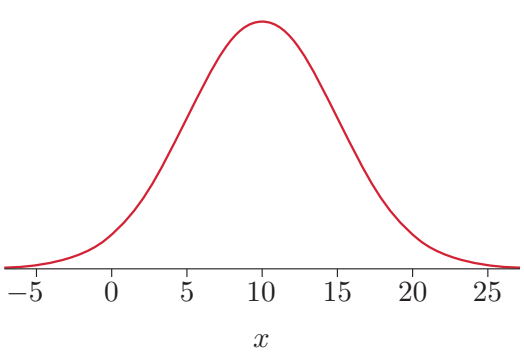


Figure 31

(b) When $\mu = -5$ and $\sigma = 2$, the normal curve is symmetric about -5 and lies between the values $-5 \pm (3 \times 2)$, that is, -11 and 1. The normal curve is illustrated in Figure 32.

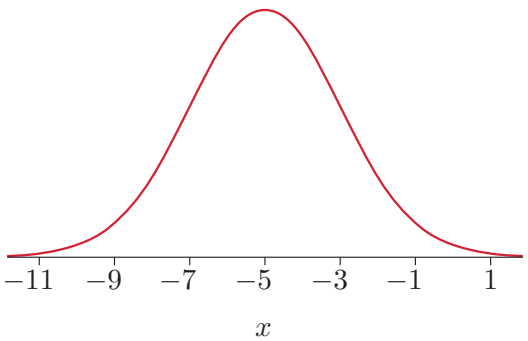


Figure 32

(c) When $\mu = 0$ and $\sigma = 1$, the normal curve is symmetric about 0 and lies between the values $0 \pm (3 \times 1)$, that is, -3 and 3. The normal curve is illustrated in Figure 33.

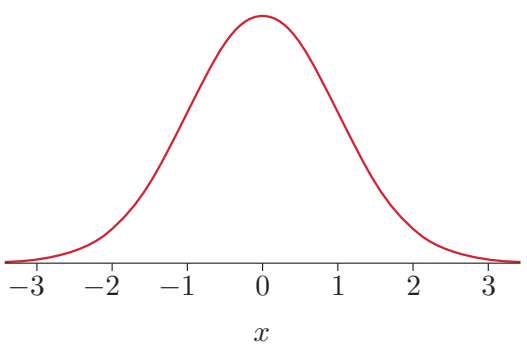


Figure 33

Solution to Activity 4

(a) The three required normal distributions (all with the same horizontal and vertical scales) are shown in Figures 34, 35 and 36; the parameter σ is 4 in each case.

(i)

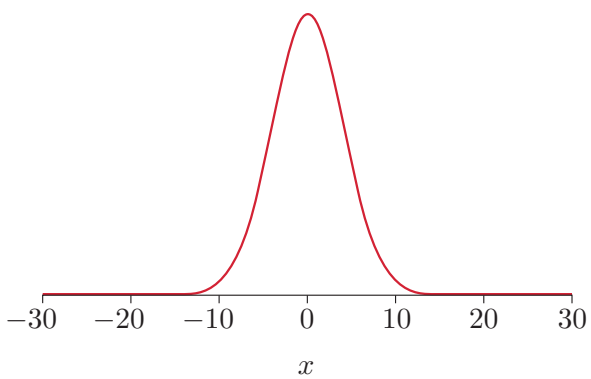


Figure 34 $\mu = 0$

(ii)

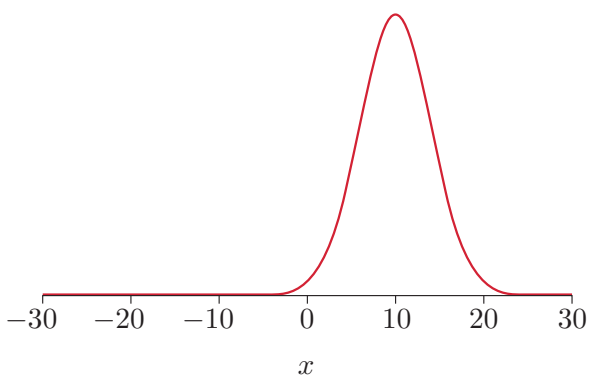


Figure 35 $\mu = 10$

(iii)

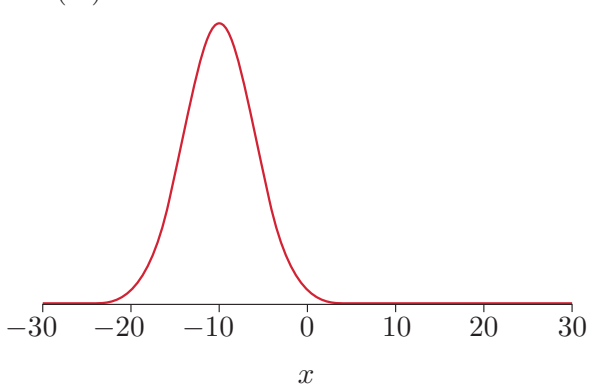


Figure 36 $\mu = -10$

(b) All normal curves are bell-shaped and symmetric about a single peak centred at μ. When μ changes, therefore, their peaks are located at different positions on the x-axis. Compared with Figure 34, which has $\mu = 0$, when μ is increased to 10 in Figure 35, the normal p.d.f. moves to the right, taking up its new location (centred at 10) but retaining its shape and spread. When μ is decreased to -10 in Figure 36, the normal p.d.f. moves to the left, taking up its new location (centred at -10) but retaining its shape and spread.

Solution to Activity 5

(a) The three required normal distributions (all with the same horizontal and vertical scales) are shown in Figures 37, 38 and 39; the parameter μ is 0 in each case.

(i)

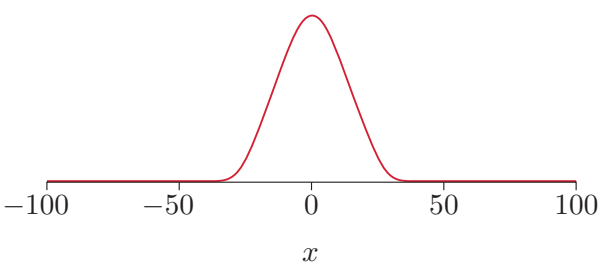


Figure 37 $\sigma = 10$

(ii)

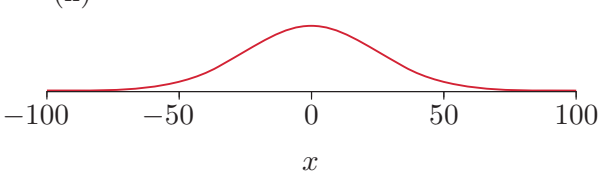


Figure 38 $\sigma = 25$

(iii)

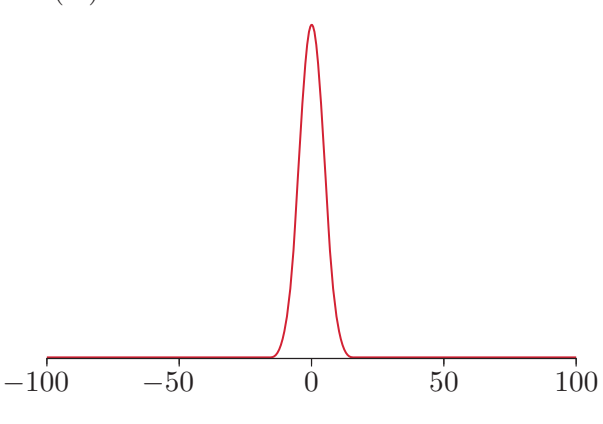


Figure 39 $\sigma = 5$

(b) All normal curves are bell-shaped and symmetric about a single peak centred at μ. Since $\mu = 0$ in all three cases, all three curves are centred at the same place, 0. Compared with Figure 37 which has $\sigma = 10$, when σ is increased to 25 in Figure 38, the normal p.d.f. flattens out, becoming more widely spread out. Conversely, when σ is decreased to 5 in Figure 39, the normal p.d.f. becomes less spread out and much 'taller'. Roughly speaking, a small value of σ produces a tall narrow 'bell', and a large value of σ produces a short wide 'bell'. (Whether a 'bell' looks short and wide or tall and narrow depends, however, on the scales used on the axes.)

Solution to Activity 6

(a) The situation described is represented in Figure 40.

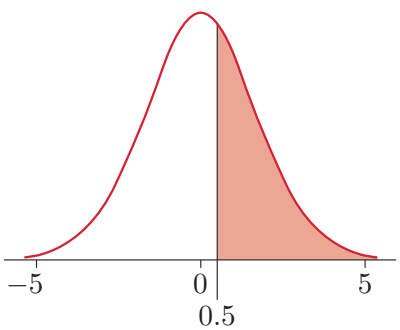


Figure 40

The shaded area in Figure 40 is given by $1 - F(0.5)$.

(b) The situation described is represented in Figure 41.

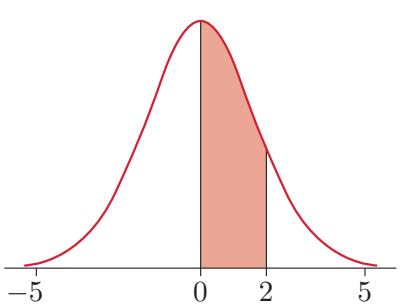


Figure 41

The shaded area in Figure 41 is given by $F(2) - F(0)$.

(c) The situation described is represented in Figure 42.

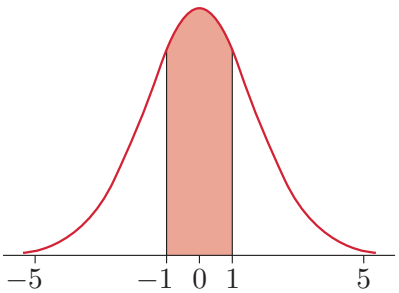


Figure 42

The shaded area in Figure 42 is equal to $F(1) - F(-1)$.

(d) The situation described is represented in Figure 43.

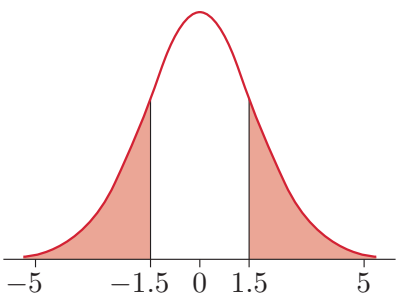


Figure 43

The shaded area on the right is equal to $1 - F(1.5)$, and the shaded area on the left is equal to $F(-1.5)$. So the whole shaded area in Figure 43 is equal to $1 - F(1.5) + F(-1.5)$ or, using the symmetry of the curve, $2 \times F(-1.5)$.

Solution to Activity 7

(a) The situation described is represented in Figure 44.

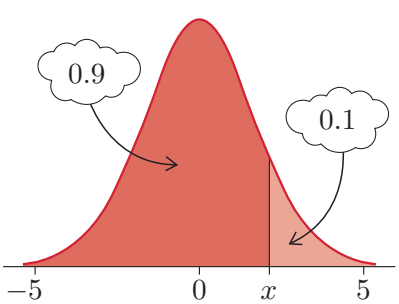


Figure 44

Since the area to the left of x is equal to 0.9, x is the 0.9-quantile of $N(0, 2.75)$: $x = q_{0.9}$.

(b) The situation described is represented in Figure 45.

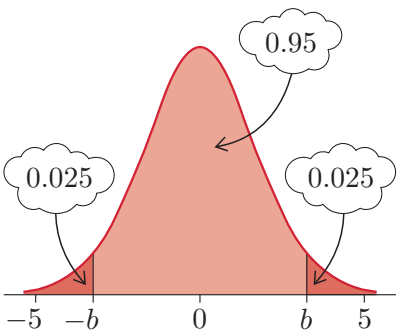


Figure 45

Since the central shaded area is equal to 0.95, the area of each of the tails is 0.025. So the area to the left of b is $0.95 + 0.025 = 0.975$. Hence b is the 0.975-quantile of $N(0, 2.75)$: $b = q_{0.975}$.

(c) The situation described is represented in Figure 46.

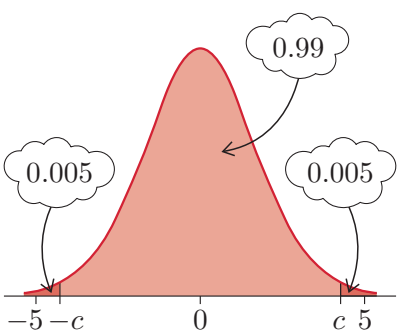


Figure 46

The area of each tail is 0.005, so the area to the left of c is 0.995 and hence $c = q_{0.995}$.

Solution to Activity 8

Using the formula for Y given in the question,

$$E(Y) = \tfrac{9}{5}\,E(X) + 32 = \left(\tfrac{9}{5} \times 452\right) + 32 = 845.6,$$

$$V(Y) = \left(\tfrac{9}{5}\right)^2 V(X) = \left(\tfrac{9}{5}\right)^2 \times 22^2 = 1568.16.$$

Given that X is normally distributed, you might have conjectured that Y is also normally distributed, with mean 845.6 and variance 1568.16.

Solution to Activity 9

Let X represent the temperature in °C, and let Y represent the temperature in °F. Then, following Activity 8,

$$Y = \tfrac{9}{5}X + 32,$$

which, rearranging (as you did in answering Exercise 11 of Unit 4), becomes

$$X = \tfrac{5}{9}(Y - 32).$$

Then

$$E(X) = \tfrac{5}{9}\,E(Y) - \tfrac{5}{9} \times 32 = \tfrac{5}{9}\,(84 - 32) \simeq 28.9$$

and

$$V(X) = \left(\tfrac{5}{9}\right)^2 V(Y) = \left(\tfrac{5}{9}\right)^2 \times 4^2 \simeq 4.9.$$

So

$$X \sim N(28.9, 4.9).$$

Solution to Activity 10

(a) If T_1 and T_2 are independent, then by Distributional Result (6), $T = T_1 + T_2$, the total time (in minutes) that Jack takes to get to work, has a normal distribution:

$$T \sim N(15 + 20, 0.5 + 9) = N(35, 9.5).$$

(b) The probability that Jack will be late for work is given by $P(T > 40)$, where $T \sim N(35, 9.5)$.

Solution to Activity 11

(a) $-Y$ is of the form $aY + b$ when $a = -1$, $b = 0$. It follows from Equation (1) that

$$E(-Y) = (-1)E(Y) + 0 = -E(Y)$$

and from Equation (2) that

$$V(-Y) = (-1)^2 V(Y) = V(Y).$$

(b) Using Equation (4),

$$E(X - Y) = E(X) + E(-Y) = E(X) - E(Y).$$

Using Equation (5),

$$V(X - Y) = V(X) + V(-Y) = V(X) + V(Y).$$

Solution to Activity 12

(a) If X_1 is a random variable representing the height of a randomly selected man and X_2 is the height of a randomly selected woman, then using Equations (7) and (8) gives

$$X_1 - X_2 \sim N(172 - 163, 19 + 13) = N(9, 32).$$

(b) The probability that the man is more than 8 cm taller than the woman is given by $P(X_1 - X_2 > 8)$, where $X_1 - X_2 \sim N(9, 32)$.

(c) The probability that the woman is taller than the man is given by $P(X_1 - X_2 < 0)$, where $X_1 - X_2 \sim N(9, 32)$.

Solution to Activity 13

(a) The probability $P(Z \leq 1.00)$ is given in the row labelled 1.0 and in the column headed 0: this gives $P(Z \leq 1.00) = 0.8413$. This probability is given by the shaded area in Figure 47.

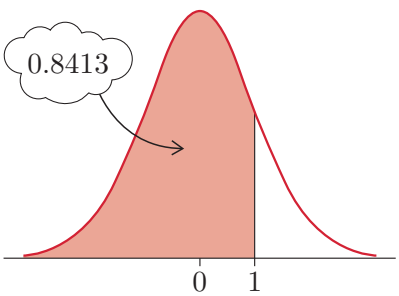


Figure 47

(b) The probability $P(Z \leq 1.96)$ is given in the row labelled 1.9 and in the column headed 6: this gives $P(Z \leq 1.96) = 0.9750$. This probability is illustrated in Figure 48.

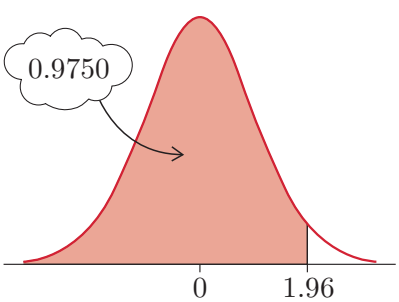


Figure 48

(c) The probability $P(Z \leq 2.25)$ is given in the row labelled 2.2 and in the column headed 5: this gives $P(Z \leq 2.25) = 0.9878$. This probability is given by the shaded area in Figure 49.

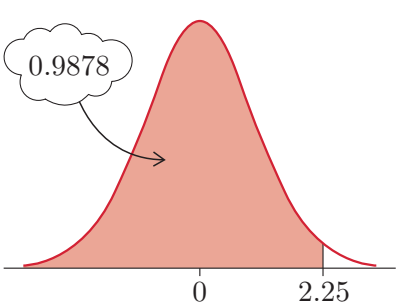


Figure 49

Solution to Activity 14

(a) The shaded area to the right in Figure 50 represents the probability $P(Z \geq 0.82)$. Since the total area under the curve is 1, this probability is equal to $1 - P(Z < 0.82)$. So, using the table,

$$P(Z \geq 0.82) = 1 - \Phi(0.82) = 1 - 0.7939 = 0.2061.$$

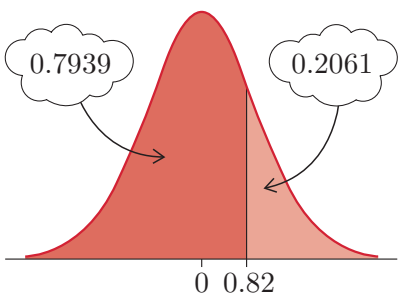


Figure 50

(b) Figure 51 shows the area required. The probability $P(0.50 \leq Z \leq 1.50)$ is given by

$$\Phi(1.50) - \Phi(0.50) = 0.9332 - 0.6915 = 0.2417.$$

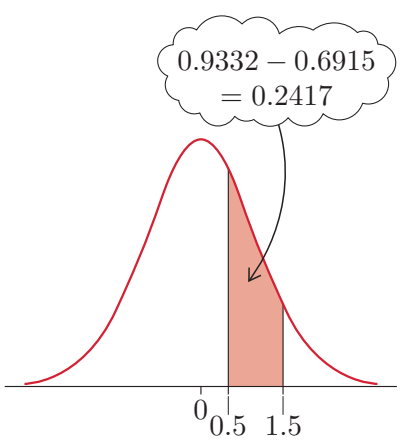


Figure 51

Solution to Activity 15

(a) Figure 52 shows the area required. Using symmetry, $P(Z < -0.66)$ is given by

$$\Phi(-0.66) = 1 - \Phi(0.66) = 1 - 0.7454 = 0.2546.$$

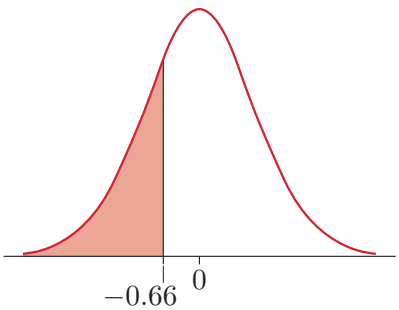


Figure 52

(b) Figure 53 shows the area required. The probability $P(-2.67 < Z < 0.33)$ is given by

$$\begin{aligned}\Phi(0.33) - \Phi(-2.67) &= \Phi(0.33) - (1 - \Phi(2.67)) \\ &= 0.6293 - (1 - 0.9962) \\ &= 0.6293 - 0.0038 = 0.6255.\end{aligned}$$

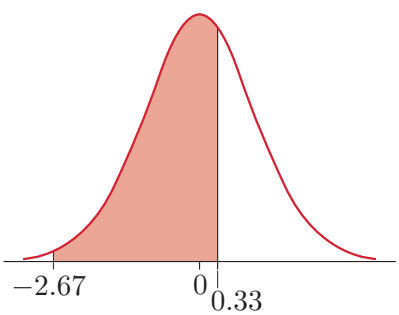


Figure 53

(c) Figure 54 shows the area required. The probability $P(-1.62 \leq Z \leq 1.62)$ is given by

$$\begin{aligned}\Phi(1.62) - \Phi(-1.62) &= \Phi(1.62) - (1 - \Phi(1.62)) = 2\Phi(1.62) - 1 \\ &= 2 \times 0.9474 - 1 = 0.8948.\end{aligned}$$

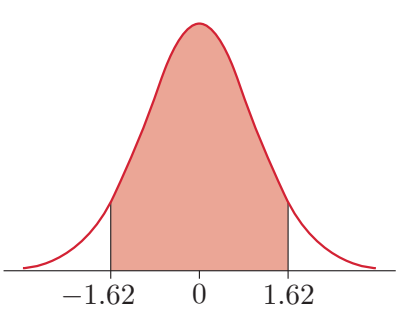


Figure 54

(d) Figure 55 shows the area required. The probability $P(|Z| \geq 2.45)$ is given by

$$\begin{aligned} P(Z \leq -2.45) + P(Z \geq 2.45) &= 2\Phi(-2.45) = 2(1 - \Phi(2.45)) \\ &= 2(1 - 0.9929) = 0.0142. \end{aligned}$$

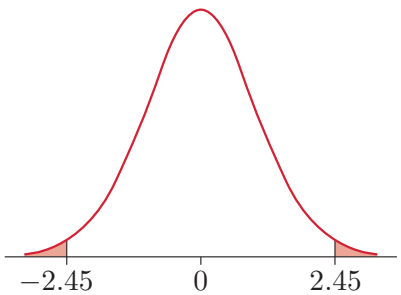


Figure 55

(e) Figure 56 shows the area required. The probability $P(-2.49 \leq Z < -0.65)$ is given by

$$\begin{aligned} P(Z < -0.65) - P(Z < -2.49) &= \Phi(-0.65) - \Phi(-2.49) \\ &= (1 - 0.7422) - (1 - 0.9936) \\ &= 0.2578 - 0.0064 = 0.2514. \end{aligned}$$

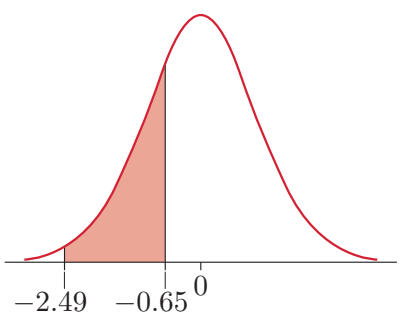


Figure 56

Solution to Activity 16

Table 6 gives the following values for the quantiles:

$$q_{0.6} = 0.2533, \quad q_{0.85} = 1.036, \quad q_{0.99} = 2.326, \quad q_{0.995} = 2.576.$$

Solution to Activity 17

Using Table 6 and the symmetry of the standard normal distribution gives the following results:

$$\begin{aligned} q_{0.4} &= -q_{0.6} = -0.2533, \\ q_{0.2} &= -q_{0.8} = -0.8416, \\ q_{0.05} &= -q_{0.95} = -1.645, \\ q_{0.01} &= -q_{0.99} = -2.326. \end{aligned}$$

Solution to Activity 18

Since $Z = (X - \mu)/\sigma$, multiplying both sides of the equation by σ gives $\sigma Z = X - \mu$, and then adding μ to both sides yields

$$X = \sigma Z + \mu.$$

To check that $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$, set $a = \sigma$ and $b = \mu$ in Distributional Result (3) to confirm that X is normally distributed with

$$E(X) = \sigma\, E(Z) + \mu = \sigma \times 0 + \mu = \mu$$

and

$$V(X) = \sigma^2\, V(Z) = \sigma^2 \times 1 = \sigma^2.$$

Solution to Activity 19

(a) According to the model, the proportion of elderly women who are shorter than 150 cm is given by the probability

$$P(H < 150) = P\left(Z < \frac{150 - 160}{6}\right) \simeq P(Z < -1.67).$$

This probability is represented by the shaded area in Figure 57. Its value is

$$\Phi(-1.67) = 1 - \Phi(1.67) = 1 - 0.9525 = 0.0475 \simeq 0.048.$$

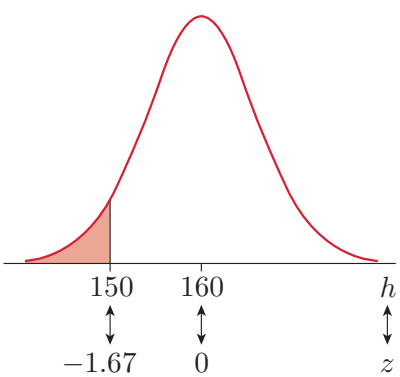


Figure 57

(b) According to the model, the proportion of elderly women who are between 155 cm and 165 cm tall is given by the probability

$$P(155 \leq H \leq 165) = P\left(\frac{155 - 160}{6} \leq Z \leq \frac{165 - 160}{6}\right)$$
$$\simeq P(-0.83 \leq Z \leq 0.83).$$

This probability is represented by the shaded area in Figure 58 (overleaf). Its value is

$$\Phi(0.83) - \Phi(-0.83) = \Phi(0.83) - (1 - \Phi(0.83)) = 2\Phi(0.83) - 1$$
$$= 2 \times 0.7967 - 1 = 0.5934 \simeq 0.593.$$

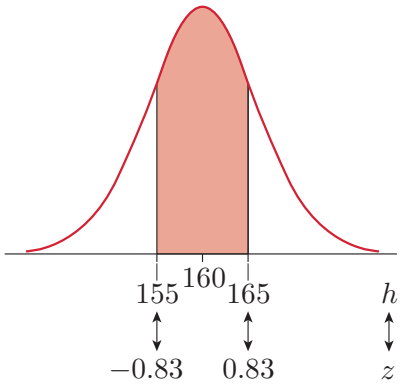


Figure 58

Solution to Activity 20

The total contents of three bags, S, has a normal distribution with mean 6009 and variance 30: $S \sim N(6009, 30)$. So the probability that the cook has less than 6 kg of sugar is given by

$$P(S < 6000) = P\left(Z < \frac{6000 - 6009}{\sqrt{30}}\right) \simeq P(Z < -1.64) = 1 - \Phi(1.64)$$
$$= 1 - 0.9495 = 0.0505 \simeq 0.051.$$

Solution to Activity 21

From the solution to Activity 10, T, the total journey time, has a normal distribution: $T \sim N(35, 9.5)$. So the probability that Jack is late for work is given by

$$P(T > 40) = P\left(Z > \frac{40 - 35}{\sqrt{9.5}}\right) \simeq P(Z > 1.62) = 1 - \Phi(1.62)$$
$$= 1 - 0.9474 = 0.0526 \simeq 0.053.$$

Solution to Activity 22

For the standard normal distribution, $q_{0.2} = -q_{0.8} = -0.8416$, $q_{0.4} = -q_{0.6} = -0.2533$, $q_{0.6} = 0.2533$ and $q_{0.8} = 0.8416$. So, assuming that scores are normally distributed with mean 100 and standard deviation 15, the corresponding quantiles for IQ scores are given by

$$q_{0.2} = 15 \times (-0.8416) + 100 \simeq 87.4,$$
$$q_{0.4} = 15 \times (-0.2533) + 100 \simeq 96.2,$$
$$q_{0.6} = 15 \times 0.2533 + 100 \simeq 103.8,$$
$$q_{0.8} = 15 \times 0.8416 + 100 \simeq 112.6.$$

These quantiles are illustrated in Figure 59.

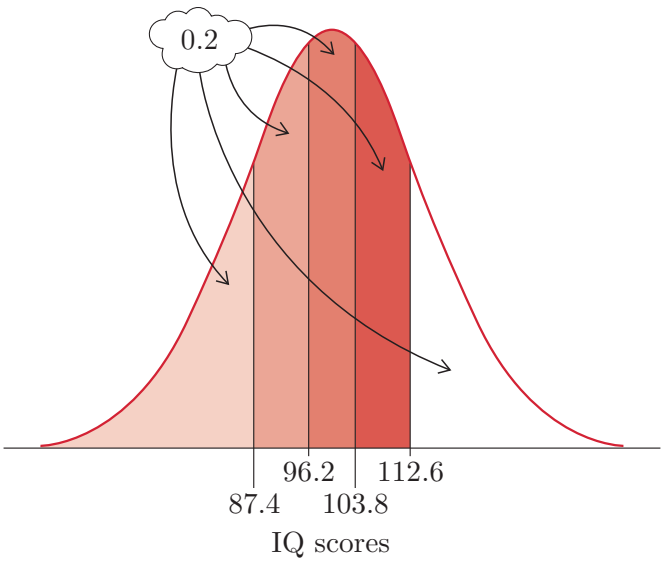


Figure 59

Solution to Activity 23

(a) The median is the 0.5-quantile of $N(160, 6^2)$, which is $6 \times 0 + 160 \simeq 160.0$ cm. The median is equal to the mean, which is in fact a general result for all normal distributions because they are all symmetric about the mean.

(b) Only 1% of elderly women are taller than the 0.99-quantile of $N(160, 6^2)$; that is, taller than $6 \times 2.326 + 160 \simeq 174.0$ cm.

(c) Only 15% of elderly women are shorter than the 0.15-quantile of $N(160, 6^2)$; that is, shorter than $6 \times (-1.036) + 160 \simeq 153.8$ cm.

Solution to Activity 24

(a) According to the model, the proportion of smokers with nicotine levels above 450 is given by

$$P(T > 450) = P\left(Z > \frac{450 - 315}{131}\right) \simeq P(Z > 1.03) = 1 - \Phi(1.03)$$
$$= 1 - 0.8485 = 0.1515 \simeq 0.152.$$

(b) According to the model, the proportion of smokers with nicotine levels below zero is given by

$$P(T < 0) = P\left(Z < \frac{0 - 315}{131}\right) \simeq P(Z < -2.40) = 1 - \Phi(2.40)$$
$$= 1 - 0.9918 = 0.0082 \simeq 0.008.$$

The model allows nearly 1% of smokers to have nicotine levels below zero. This is large enough to question the adequacy of the normal model.

(c) The quartiles of the distribution are given by

$$q_{0.75} = 131 \times 0.6745 + 315 \simeq 403.36,$$
$$q_{0.25} = 131 \times (-0.6745) + 315 \simeq 226.64.$$

So the interquartile range of nicotine levels is $403.36 - 226.64 \simeq 176.7$.

(d) Only 4% of smokers have nicotine levels above the 0.96-quantile, which is given by $131 \times 1.751 + 315 \simeq 544.4$.

Solution to Activity 25

(a) For a random sample, the components in the sample total T_n are identically distributed. In particular, their means are equal, and applying Equation (4),

$$\begin{aligned} E(T_n) &= E(X_1 + X_2 + \cdots + X_n) \\ &= E(X_1) + E(X_2) + \cdots + E(X_n) = n\mu. \end{aligned}$$

(b) Since $\overline{X}_n = T_n/n$, applying Equation (1) with $a = 1/n$ and $b = 0$ gives

$$E(\overline{X}_n) = E\left(\frac{1}{n}T_n\right) = \frac{1}{n}E(T_n)$$

and hence

$$E(\overline{X}_n) = \frac{1}{n}n\mu = \mu.$$

Solution to Activity 26

(a) For a random sample, the components in the sample total T_n are identically distributed. In particular, their variances are equal. They are also independent, so applying Equation (5),

$$\begin{aligned} V(T_n) &= V(X_1 + X_2 + \cdots + X_n) \\ &= V(X_1) + V(X_2) + \cdots + V(X_n) = n\sigma^2. \end{aligned}$$

(b) Since $\overline{X}_n = T_n/n$, applying Equation (2) with $a = 1/n$ and $b = 0$ gives

$$V(\overline{X}_n) = V\left(\frac{1}{n}T_n\right) = \frac{1}{n^2}V(T_n)$$

and hence

$$V(\overline{X}_n) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

That is, the variance of the sample mean is equal to the population variance divided by the sample size.

Solution to Activity 27

(a) If X represents the contents of one bag of sugar and T_5 represents the total contents of five bags of sugar, then

$$E(T_5) = 5E(X) = 5015$$

and

$$V(T_5) = 5V(X) = 5 \times 4 = 20.$$

So the mean total contents of five bags is 5015 g, and the standard deviation of the total contents is $\sqrt{20}$ g, or approximately 4.5 g.

(b) The expected value of $\overline{X}_5$ is 1003 g. The variance is given by

$$V(\overline{X}_5) = \tfrac{1}{5}V(X) = \tfrac{1}{5} \times 4 = 0.8,$$

so the standard deviation of $\overline{X}_5$ is $\sqrt{0.8}$ g, or approximately 0.9 g.

Solution to Activity 28

By the Central Limit Theorem, Distributional Result (21),

$$\overline{X}_{40} \approx N(30, 1^2/40) = N(30, 0.025).$$

So the required probability is given by

$$P(\overline{X}_{40} < 29.75) \simeq P\left(Z < \frac{29.75 - 30}{\sqrt{0.025}}\right) \simeq P(Z < -1.58) = 1 - \Phi(1.58)$$
$$= 1 - 0.9429 = 0.0571 \simeq 0.057.$$

Solution to Activity 29

The sample mean $\overline{X}_{100}$ of a random sample from the geometric distribution has

$$E(\overline{X}_{100}) = \mu = \frac{1}{p}, \quad V(\overline{X}_{100}) = \frac{\sigma^2}{n} = \frac{1-p}{100\,p^2},$$

so

$$\overline{X}_{100} \approx N\left(\frac{1}{p}, \frac{1-p}{100\,p^2}\right).$$

Solution to Activity 30

(a) If X represents the error in an individual transaction, then $X \sim U(-\frac{1}{2}, \frac{1}{2})$. Using the hint given in the question, in this case, X has mean

$$E(X) = \frac{-\frac{1}{2} + \frac{1}{2}}{2} = 0$$

and variance

$$V(X) = \frac{\left(\frac{1}{2} - \left(-\frac{1}{2}\right)\right)^2}{12} = \frac{1}{12}.$$

The total error accumulated in 400 transactions is given by the sum

$$T_{400} = X_1 + X_2 + \cdots + X_{400}.$$

By the corollary to the Central Limit Theorem, the distribution of T_{400} is approximately normal with mean

$$E(T_{400}) = 400 \times 0 = 0$$

and variance

$$V(T_{400}) = 400 \times \frac{1}{12} = \frac{100}{3}.$$

That is, $T_{400} \approx N(0, 100/3)$.

(b) The probability that the error in his estimate is less than £10 is

$$P(-10 < T_{400} < 10) \simeq P\left(\frac{-10-0}{\sqrt{100/3}} < Z < \frac{10-0}{\sqrt{100/3}}\right)$$
$$\simeq P(-1.73 < Z < 1.73) = \Phi(1.73) - \Phi(-1.73)$$
$$= \Phi(1.73) - (1 - \Phi(1.73)) = 2\Phi(1.73) - 1$$
$$= 2 \times 0.9582 - 1 = 0.9164 \simeq 0.916.$$

Solution to Activity 31

From the solution to Activity 27, the sample mean $\overline{X}_5$ has mean 1003 g and standard deviation $\sqrt{0.8}$ g. So the probability that the mean contents of a sample of five bags is less than 1 kg is

$$P(\overline{X}_5 < 1000) = P\left(Z < \frac{1000-1003}{\sqrt{0.8}}\right) \simeq P(Z < -3.35) = 1 - \Phi(3.35)$$
$$= 1 - 0.9996 = 0.0004.$$

(Rounding to 3 d.p. would give 0.000; in such cases, it seems appropriate to show the fourth decimal place too if, as here, it's non-zero.)

Solutions to exercises

Solution to Exercise 1

In each case, the shape of the normal p.d.f. remains the same.

(a) Compared with the distribution of X, the mean of Y is smaller (0.5 compared with 1) and the variance of Y is smaller (also 0.5 compared with 1). Hence the p.d.f. associated with Y will be moved to the left and will be less spread out ('taller') than the p.d.f. associated with X.

(b) Compared with the distribution of X, the mean of Y is the same (-1 compared with -1) and the variance of Y is smaller (1 compared with 3). Hence the p.d.f. associated with Y will be located in the same place but will be less spread out ('taller') than the p.d.f. associated with X.

(c) Compared with the distribution of X, the mean of Y is smaller (-1 compared with 0) and the variance of Y is larger (100 compared with 10). Hence the p.d.f. associated with Y will be moved to the left and will be flattened out compared with the p.d.f. associated with X.

(d) Compared with the distribution of X, the mean of Y is larger (0 compared with -5) and the variance of Y is the same (1 compared with 1). Hence the p.d.f. associated with Y will be moved to the right compared with the p.d.f. associated with X, and will be otherwise the same.

Solution to Exercise 2

(a) We have $\mu = 100$ and $\sigma = 15$, so the normal curve should be bell-shaped and symmetric about 100, with most of the distribution between the values $100 \pm (3 \times 15)$, that is, between 55 and 145. A sketch of the normal curve is given in Figure 60.

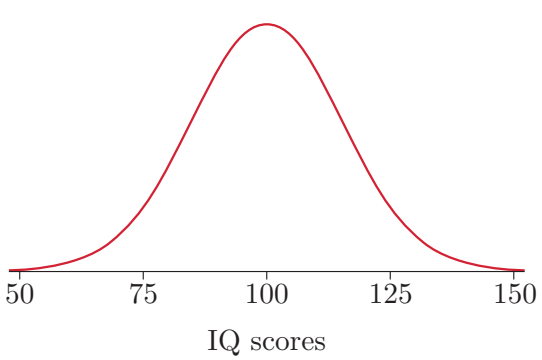


Figure 60

(b) The probability of a score between 110 and 130 is indicated in Figure 61 (overleaf).

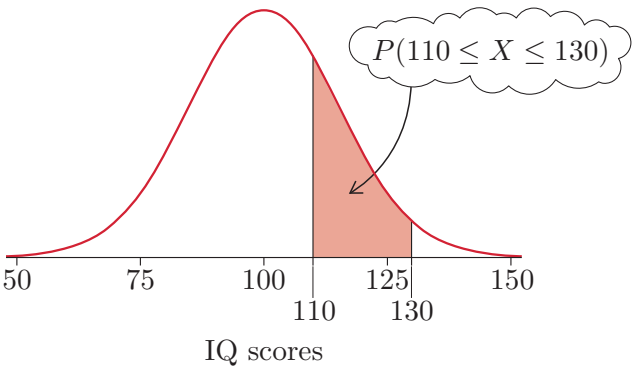


Figure 61

This area is equal to the area to the left of 130 minus the area to the left of 110; that is,

$$P(110 \leq X \leq 130) = F(130) - F(110).$$

(c) Since 5% of scores are greater than x,

$$P(X > x) = 0.05.$$

This means that $F(x) = 1 - 0.05 = 0.95$, so $x = q_{0.95}$, the 0.95-quantile of $N(100, 15^2)$.

Solution to Exercise 3

Since $Y = 3X + 2$ is a linear function of X, it also has a normal distribution. Its mean is given by

$$E(Y) = 3E(X) + 2 = 3 \times 10 + 2 = 32.$$

The variance is

$$V(Y) = 3^2 V(X) = 9 \times 5 = 45.$$

That is, $Y \sim N(32, 45)$.

Solution to Exercise 4

Since X_1, X_2 and X_3 are independent normal random variables, S also has a normal distribution. Its mean is given by

$$E(S) = E(X_1) + E(X_2) + E(X_3) = -10 + 15 - 2 = 3.$$

The variance is

$$V(S) = V(X_1) + V(X_2) + V(X_3) = 2 + 12 + 5 = 19.$$

That is, $S \sim N(3, 19)$.

Solution to Exercise 5

Since X and Y are independent normal random variables, $X - Y$ also has a normal distribution. Its mean is given by

$$E(X - Y) = E(X) - E(Y) = -5 - 10 = -15.$$

The variance is

$$V(X - Y) = V(X) + V(Y) = 4 + 5 = 9.$$

That is, $X - Y \sim N(-15, 9)$.

Solution to Exercise 6

(a) Figure 62 shows the area required. Using symmetry, $P(Z > -0.42)$ is given by

$$1 - \Phi(-0.42) = 1 - (1 - \Phi(0.42)) = \Phi(0.42) = 0.6628 \simeq 0.663.$$

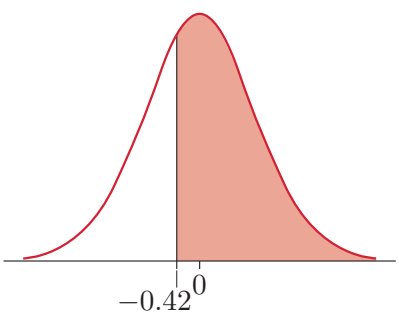


Figure 62

(b) Figure 63 shows the area required. The probability $P(0.17 < Z < 1.17)$ is given by

$$\Phi(1.17) - \Phi(0.17) = 0.8790 - 0.5675 = 0.3115 \simeq 0.312.$$

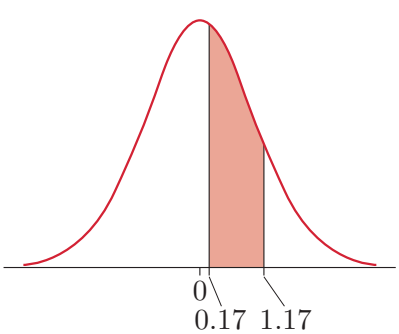


Figure 63

(c) Figure 64 (overleaf) shows the area required. The probability $P(|Z| \leq 2.25) = P(-2.25 \leq Z \leq 2.25)$ is given by

$$\begin{aligned}\Phi(2.25) - \Phi(-2.25) &= \Phi(2.25) - (1 - \Phi(2.25)) = 2\Phi(2.25) - 1 \\ &= 2 \times 0.9878 - 1 = 0.9756 \simeq 0.976.\end{aligned}$$

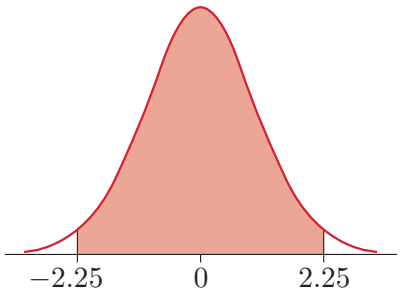


Figure 64

(d) Figure 65 shows the area required. The probability $P(|Z| \geq 1.75)$ is given by

$$P(Z \leq -1.75) + P(Z \geq 1.75) = 2\Phi(-1.75) = 2(1 - \Phi(1.75))$$
$$= 2(1 - 0.9599) = 0.0802 \simeq 0.080.$$

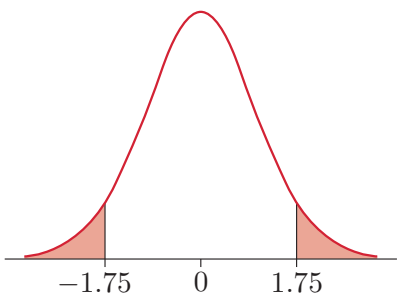


Figure 65

Solution to Exercise 7

Table 6 gives the following values: $q_{0.998} = 2.878$, $q_{0.55} = 0.1257$. Using Table 6 and the symmetry of the standard normal distribution gives the following results:

$$q_{0.35} = -q_{0.65} = -0.3853,$$
$$q_{0.1} = -q_{0.9} = -1.282,$$
$$q_{0.001} = -q_{0.999} = -3.090.$$

Solution to Exercise 8

(a) If H is a random variable representing the height of a man in the population, then according to the model, the proportion of men who are shorter than 160 cm is given by the probability

$$P(H < 160) = P\left(Z < \frac{160 - 173}{\sqrt{40}}\right) \simeq P(Z < -2.06) = 1 - \Phi(2.06)$$
$$= 1 - 0.9803 = 0.0197 \simeq 0.020.$$

(b) According to the model, the proportion of men in the population who are between 170 cm and 180 cm tall is given by the probability

$$\begin{aligned} P(170 \leq H \leq 180) &= P\left(\frac{170-173}{\sqrt{40}} \leq Z \leq \frac{180-173}{\sqrt{40}}\right) \\ &\simeq P(-0.47 \leq Z \leq 1.11) \\ &= \Phi(1.11) - \Phi(-0.47) = \Phi(1.11) - (1 - \Phi(0.47)) \\ &= 0.8665 - (1 - 0.6808) = 0.5473 \simeq 0.547. \end{aligned}$$

(c) Only 2% of men in the population are taller than the 0.98-quantile of $N(173, 40)$, that is, taller than $\sqrt{40} \times 2.054 + 173 \simeq 186.0$ cm.

(d) The quartiles of the distribution are given by

$$\begin{aligned} q_{0.75} &= \sqrt{40} \times 0.6745 + 173 \simeq 177.27, \\ q_{0.25} &= \sqrt{40} \times (-0.6745) + 173 \simeq 168.73. \end{aligned}$$

So the interquartile range of heights is
$177.27 - 168.73 = 8.54 \simeq 8.5$ cm.

Solution to Exercise 9

If $X_1, X_2, \ldots, X_{10}$ represent the contents in grams of ten randomly selected bags of flour, then by Distributional Result (6), T, the total contents of the ten bags, has a normal distribution:

$$T \sim N(10 \times 1501, 10 \times 2.5) = N(15\,010, 25).$$

So the probability that the total contents is less than 15 kg is given by

$$\begin{aligned} P(T < 15\,000) &= P\left(Z < \frac{15\,000 - 15\,010}{5}\right) = P(Z < -2) = 1 - \Phi(2) \\ &= 1 - 0.9772 = 0.0228 \simeq 0.023. \end{aligned}$$

Solution to Exercise 10

If X represents the weight of a single item, then the distribution of T_{30}, the total weight of 30 items, is, by the corollary to the Central Limit Theorem, approximately normal with mean and variance given by

$$\begin{aligned} E(T_{30}) &= 30 \times 50 = 1500, \\ V(T_{30}) &= 30 \times 100 = 3000. \end{aligned}$$

That is, $T_{30} \approx N(1500, 3000)$. Hence the probability that the total weight of 30 items is less than 1600 kg is given by

$$\begin{aligned} P(T_{30} < 1600) &\simeq P\left(Z < \frac{1600 - 1500}{\sqrt{3000}}\right) \simeq P(Z < 1.83) \\ &= \Phi(1.83) = 0.9664 \simeq 0.966. \end{aligned}$$

Solution to Exercise 11

(a) If X represents the contents of one bottle of orange juice and T_3 represents the total contents of three randomly selected bottles, then

$$E(T_3) = 3E(X) = 3012$$

and

$$V(T_3) = 3V(X) = 3 \times 25 = 75.$$

So $T_3 \sim N(3012, 75)$ and the probability that the total contents of three bottles will be less than 3 litres is given by

$$P(T_3 < 3000) = P\left(Z < \frac{3000 - 3012}{\sqrt{75}}\right) \simeq P(Z < -1.39)$$
$$= 1 - \Phi(1.39) = 1 - 0.9177 = 0.0823 \simeq 0.082.$$

(b) If $\overline{X}_4$ represents the mean contents of four randomly selected bottles, then the expected value of $\overline{X}_4$ is 1004 and the variance of $\overline{X}_4$ is given by

$$V(\overline{X}_4) = \tfrac{1}{4}V(X) = \tfrac{1}{4} \times 25 = 6.25,$$

and hence $\overline{X}_4 \sim N(1004, 6.25)$. So the probability that the mean contents of four bottles will be less than 1 litre is given by

$$P(\overline{X}_4 < 1000) = P\left(Z < \frac{1000 - 1004}{\sqrt{6.25}}\right) = P(Z < -1.6)$$
$$= 1 - \Phi(1.6) = 1 - 0.9452 = 0.0548 \simeq 0.055.$$

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 4 top: © Peter Junaidy / Dreamstime.com

Page 4 bottom: © 2003–2016 Andy Sweet / Stravaiging around Scotland

Page 14: Taken from: http://www.hawking.org.uk/images.html

Page 15: mearicon/www.123rf.com

Page 16: Taken from:
http://www.essentialchemicalindustry.org/processes/cracking-isomerisation-and-reforming.html

Page 17: Tobkatrina/www.123rf.com

Page 18 top: Anek Suwannaphoom/www.123rf.com

Page 18 bottom: Samart Boonywang/www.123rf.com

Page 19: © Iofoto/Dreamstime.com

Page 23: © KC O Connor/eacgs.com. Taken from www.eacgs.com. Reproduced with permission

Page 26: © Xload/Deposit Photos.com

Page 30: Cathy Keifer/www.123rf.com

Page 34: AmbientIdeas/www.istockphoto.com

Page 36 top: Image Copyright © Tony Danbury

Page 41: Ruud Morijn/www.123rf.com

Page 45: Celebration of the 275th anniversary of the normal distribution at Dept of Statistics, University of Chicago. Copyright © 2008 University of Chicago.

Page 46: nattanan726/www.istockphoto.com

Page 48 top: © SMS Market Research

Page 48 bottom: © Seankate/Dreamstime.com

Page 50 top: © Catalin205/Dreamstime.com

Page 50 bottom: © Mlan61/Dreamstime.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.